Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2013)
75, *Part* 3, *pp.*

Comparing distributions by using dependent normalized random-measure mixtures

J. E. Griffin,

University of Kent, Canterbury, UK

M. Kolossiatis

Cyprus University of Technology, Limassol, Cyprus

and M. F. J. Steel

University of Warwick, Coventry, UK

[Received December 2010. Final revision September 2012]

Summary. A methodology for the simultaneous Bayesian non-parametric modelling of several distributions is developed. Our approach uses normalized random measures with independent increments and builds dependence through the superposition of shared processes. The properties of the prior are described and the modelling possibilities of this framework are explored in detail. Efficient slice sampling methods are developed for inference. Various posterior summaries are introduced which allow better understanding of the differences between distributions. The methods are illustrated on simulated data and examples from survival analysis and stochastic frontier analysis.

Keywords: Bayesian non-parametrics; Dependent distributions; Dirichlet process; Normalized generalized gamma process; Slice sampling; Utility function

1. Introduction

This paper considers the non-parametric modelling of data divided into different groups and the comparison of their distributions. For example, we may observe the results of different medical treatments or the performance of firms with different management structures. Statistical analysis will often concentrate on inference about the differences in the distributions. Analysis of variance (ANOVA) focuses on differences between means for different groups and links these to the effects of each factor. However, differences between groups may not be well modelled by restricting attention to location. For example, if there are distinct subpopulations within the observations then each group may contain different proportions of each subpopulation and a full summary of the differences would involve identifying parts of the support on which the two distributions place substantially different masses. We implement a fully Bayesian analysis by firstly placing a prior on the distributions and secondly defining a decision analysis which reports where the distributions are similar or substantially different.

Address for correspondence: M. F. J. Steel, Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK.
E-mail: M.F.Steel@stats.warwick.ac.uk

 1369–7412/13/75000

We use a Bayesian non-parametric mixture model approach to understand the differences between the distributions. Let $F_1, F_2, \ldots, F_q$ be the (continuous) distributions of the observations for q different groups; then an infinite mixture model assumes that the density for the gth group is

$$f_g = \int k(\cdot|\theta)\, \mathrm{d}G_g(\theta)$$

where $k(\cdot|\theta)$ is a density parameterized by θ and G_g is a discrete random-probability measure. Since the measure is discrete, it can be represented as

$$G_g = \sum_{i=1}^{\infty} w_{g,i}\delta_{\theta_{g,i}}$$

where δ_x is the Dirac delta function that places mass 1 at x and $\theta_{g,1}, \theta_{g,2}, \ldots$ and $w_{g,1}, w_{g,2}, \ldots$ are infinite sequences of random variables for which $\Sigma_{i=1}^{\infty} w_{g,i} = 1$ and $w_{g,i} > 0$ for all g and i. It follows that the mixture model for group g can be written as

$$f_g = \sum_{i=1}^{\infty} w_{g,i}\, k(\cdot|\theta_{g,i}) \tag{1}$$

or, alternatively, the model can be represented hierarchically for an observation $y_{g,j}$ drawn from F_g as follows:

$$y_{g,j} \sim k(\cdot|\theta_{g,s_{g,j}}),$$
$$p(s_{g,j} = i) = w_{g,i}$$

where $s_{g,j}$ is an allocation variable indicating to which component distribution $k(\cdot|\theta)$ the jth observation in group g is allocated. The groups will often be formed by all possible combinations of some categorical covariates and we shall denote those covariates by z_g for the gth group. This is a very general model which encompasses many previously proposed models. The ANOVA dependent Dirichlet process model of De Iorio *et al.* (2004) assumes that $w_{g,i} = w_i$ and the density k is a normal distribution with mean $z_g^{\mathrm{T}}\beta_i$, where β_i is a vector of parameters, and variance σ^2. This allows the means of the different components to change with covariates.

A popular approach allows the weights to depend on covariates and sets $\theta_{g,i} = \theta_i$ so that the location of the components is fixed across each group. A finite mixture of normals model along these lines was proposed by Rodriguez *et al.* (2009) who allowed the component weights to depend on covariates. Alternatively, the weights can be modelled through combinations of random variables, which encourages correlation between the random distributions. The matrix stick breaking process of Dunson *et al.* (2008) focuses on groups which are formed by the combination of two factors, say $z_{g,1}$ and $z_{g,2}$, each with a finite number of levels. They defined the weights by using the stick breaking construction

$$w_{g,i} = V_{z_{g,1},i} U_{z_{g,2},i} \prod_{l<i} (1 - V_{z_{g,1},l} U_{z_{g,2},l})$$

where $U_{j,i}$ and $V_{j,i}$ are independent, beta-distributed random variables. Müller *et al.* (2004) assumed that

$$f_g = \psi \sum_{i=1}^{\infty} w^*_{g,i}\, k(\cdot|\theta^*_{g,i}) + (1-\psi) \sum_{i=1}^{\infty} w_i k(\cdot|\theta_i)$$

where $0 \leqslant \psi \leqslant 1$. The distribution of the gth group is a mixture of a common component shared by all groups and an idiosyncratic component. The parameter ψ is the weight that is placed on the idiosyncratic component and so affects the correlation between distributions.

The hierarchical Dirichlet process (Teh *et al.*, 2006) assumes, in its simplest form, that

$$G_g|G_0 \overset{\text{IID}}{\sim} \text{DP}(MG_0), \qquad g=1,\ldots,q, \quad G_0 \sim \text{DP}(M_0 H), \tag{2}$$

where DP(MH) indicates a Dirichlet process with mass parameter $M>0$ and centring (or base) distribution H. The distributions are exchangeable and this structure allows clusters to be shared by different groups (owing to the discrete nature of the Dirichlet process at both levels). If the hierarchical Dirichlet process is used as the mixture distribution in the mixture models then we have something of the form of model (1). Teh *et al.* (2006) derived the stick breaking construction for $w_{g,1}, w_{g,2}, w_{g,3}, \ldots$. The model can be extended to more levels of hierarchy in the standard way. This model assumes that distributions are exchangeable at some level. In contrast, the present paper will mostly concentrate on the problem where groups are defined by covariates. There is normally no natural nesting in these settings, so hierarchical models will then not be appropriate. Note also that we shall focus on categorical covariates, which naturally lead to a finite number of groups.

We propose to use a normalized superposition of random measures to induce dependence. This general framework leads to dependence structures that can be fairly easily controlled through the mass parameters of the underlying component measures and extends naturally to any number of groups. In fact, we can use this framework to model separately the mass shared by any subset of the groups or we can use simpler settings, depending on the flexibility of the dependence structure that we want to assume. We can formally conduct model selection by using point mass priors on the mass parameters corresponding to the components. Alternatively, we use shrinkage priors for the mass parameters to ensure consistent priors across different levels of model complexity. For posterior inference, we propose slice sampling Markov chain Monte Carlo (MCMC) methods, used in combination with a split–merge move. We also introduce ways of summarizing the differences between the non-parametric distributions for each group, based on decision theoretic ideas.

The paper is organized as follows. Section 2 describes the construction of random probability measures by normalization and our proposed framework for modelling dependence by using normalized random measures, Section 3 describes efficient MCMC sampling methods for inference, Section 4 discusses a decision theoretic approach to comparing distributions, Section 5 analyses simulated data and presents real data applications to survival analysis and stochastic frontier modelling, whereas Section 6 concludes.

2. Introducing dependence in normalized random measures

2.1. General framework

Normalized random measures with independent increments (known in the literature as 'NRMIs') are a class of non-parametric priors for a random probability measure G constructed by normalizing a positive random measure $\tilde{G}$ with independent increments and support Ω (usually, a subset of $\mathbb{R}^m$). Thus, for any measurable set $B \in \Omega$ we define

$$G(B) = \tilde{G}(B)/\tilde{G}(\Omega).$$

Throughout the paper we shall use G to represent the normalized version of a random measure $\tilde{G}$. Generally, we shall concentrate on random measures which only contain jumps and write

$$\tilde{G} = \sum_{i=1}^{\infty} J_i \delta_{\theta_i},$$

where θ_i are independent and identically distributed with some distribution H and $J_1, J_2, J_3, \ldots$ are jumps of a Lévy process with Lévy density $\zeta(\cdot)$. The process is well defined if $0 < \tilde{G}(\Omega) < \infty$

almost surely which happens if $\int \zeta(x)\,\mathrm{d}x$ is infinite. The NRMI can be employed as the prior of the mixing measure G in an infinite mixture model $f(y)=\int k(y|\theta)\,\mathrm{d}G(\theta)$ to define an NRMI mixture. This class of processes and their use in mixture models was studied in general by James *et al.* (2009). We focus on homogeneous NRMIs, which implies *a priori* independence between the jumps and the locations. James *et al.* (2009) showed that these processes can be defined on arbitrary spaces. Several previously proposed processes fall within this class. The Dirichlet process (Ferguson, 1973) occurs if $\tilde{G}$ is a gamma process, for which

$$\zeta(x)=Mx^{-1}\exp(-x), \qquad M>0.$$

This can be generalized to the normalized generalized gamma (NGG) process (Lijoi *et al.*, 2007) which is constructed by normalizing a generalized gamma process (Brix, 1999) for which

$$\zeta(x)=\frac{M}{\Gamma(1-a)}x^{-1-a}\exp(-\lambda x), \qquad M>0, \quad 0<a<1, \quad \lambda\geqslant 0. \tag{3}$$

This process tends to the Dirichlet process as $a\to 0$ and $\lambda=1$. The normalized inverse Gaussian process (Lijoi *et al.*, 2005) occurs if $a=0.5$ and $\lambda=1$. Another special case is the normalized stable process of Kingman (1975), which corresponds to $\lambda=0$.

Dependence between two distributions G_1 and G_2 is introduced through the unnormalized random measures $\tilde{G}_1$ and $\tilde{G}_2$. Intuitively, it is clear that the dependence between G_1 and G_2 will grow as the dependence between $\tilde{G}_1$ and $\tilde{G}_2$ grows. A similar approach for constructing processes of random probability measures over time was discussed by Griffin (2011).

Suppose that we have q groups; then the random measures are defined in the following way. Firstly, we define p underlying random measures $\tilde{G}_1^*, \tilde{G}_2^*, \ldots, \tilde{G}_p^*$ such that

$$\tilde{G}_j^*=\sum_{i=1}^{\infty} J_{j,i}\delta_{\theta_{j,i}}, \qquad j=1,\ldots,p,$$

where $\theta_{j,i}$ are independent and identically distributed with some distribution H and $J_j=(J_{j,1},J_{j,2},J_{j,3},\ldots)$ are the jumps of a Lévy process with Lévy density $\zeta_j^*(\cdot)$. It is assumed that $J_1, J_2, \ldots, J_p$ are independent. Defining $\tilde{G}^*=(\tilde{G}_1^*,\tilde{G}_2^*,\ldots,\tilde{G}_p^*)^{\mathrm{T}}$, the random measures in the vector $\tilde{G}=(\tilde{G}_1,\tilde{G}_2,\ldots,\tilde{G}_q)^{\mathrm{T}}$ will be formed as

$$\tilde{G}=D\tilde{G}^*,$$

where D is a $(q\times p)$-dimensional selection matrix (i.e. a matrix with only 0s and 1s as elements). Then $\tilde{G}_j$ is a Lévy process and the Lévy density of $\tilde{G}_j$ is $\zeta_j(x)=D_{j\cdot}\zeta^*(x)$ where $D_{j\cdot}$ is the jth row of D and $\zeta^*(x)=(\zeta_1^*(x),\ldots,\zeta_p^*(x))^{\mathrm{T}}$. In particular, we take $\zeta_h^*(x)=M_h\,\eta(x)$ so that $\zeta_j(x)=B_j\,\eta(x)$ where $B_j=\sum_{k=1}^{p}D_{jk}M_k$ for $j=1,2,\ldots,q$. When we normalize, we obtain

$$G=WG^*, \tag{4}$$

where $G=(G_1,\ldots,G_q)^{\mathrm{T}}$, $G^*=(G_1^*,\ldots,G_p^*)^{\mathrm{T}}$ with

$$G_j^*=\tilde{G}_j^*/\tilde{G}_j^*(\Omega)$$

and W is a $q\times p$ matrix with elements

$$W_{ij}=D_{ij}\tilde{G}_j^*(\Omega)\Big/\sum_{k=1}^{p}D_{ik}\tilde{G}_k^*(\Omega).$$

Therefore, the distribution for each group is a mixture of $G_1^*, G_2^*, \ldots, G_p^*$ where the weights for the ith group are given by the ith row of W. This process will be denoted generally as a *correlated NRMI* process or CNRMI$(M,H,D;\eta)$ where $M=(M_1,\ldots,M_p)$. Often, we shall choose a specific functional form for η so that the marginal processes $G_1,\ldots,G_q$ correspond to a known

process (e.g. a Dirichlet process). We shall consider two possibilities: a *correlated Dirichlet process* $\mathrm{CDP}(M, H, D)$ where $\eta(x) = x^{-1}\exp(-x)$ and the marginal processes are Dirichlet processes and a *correlated normalized generalized gamma process* $\mathrm{CNGG}(M, H, D; a, \lambda)$ where $\eta(x) = x^{-1-a}\exp(-\lambda x)/\Gamma(1-a)$ and the marginal processes are NGG processes. The mixture form for $G_1, G_2, \ldots, G_q$ is an important difference from the hierarchical Dirichlet process, which is a framework that leads to all atoms being shared by all distributions and assumes that all distributions are *a priori* equally correlated.

If we use $G_1, G_2, \ldots, G_q$ as mixing measures for q mixture models, the density of an observation $y \in \mathcal{Y}$ in the ith group is now given by

$$f_i(y) = \int k(y|\theta)\,\mathrm{d}G_i(\theta).$$

Then we can write

$$f_i = \tilde{f}_i / \tilde{F}_i(\mathcal{Y}),$$

where $\tilde{f}_i(y) = \int k(y|\theta)\,\mathrm{d}\tilde{G}_i(\theta)$ and $\tilde{F}_i(A) = \int_A \tilde{f}_i(y)\,\mathrm{d}y$. Now, $\tilde{f}_i$ expresses an unnormalized density in terms of basis functions (where the density $k(\cdot)$ describes the basis functions) and so f_i is a normalized basis function model.

2.2. Dependence between distributions

A natural measure of the dependence between two distributions is the correlation between $G_i(B)$ and $G_j(B)$ where B is a measurable set. Using the construction in this paper, this correlation does not depend on B and so can be used as a single measure of dependence between distributions, which we denote by $\mathrm{corr}(G_i, G_j)$. The following results present an expression for the correlation, using a particular form of the framework described above for $q = 2$, $p = 3$ and

$$D = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

This is a simple, yet illustrative, example.

Theorem 1. Suppose that $\tilde{G}_1 = \tilde{G}_1^* + \tilde{G}_3^*$ and $\tilde{G}_2 = \tilde{G}_2^* + \tilde{G}_3^*$ where the Lévy measure of $\tilde{G}_k^*$ is $M_k\,\eta(x)$ and define

$$L_\eta(v) = \int_0^\infty \{1 - \exp(-vx)\}\,\eta(x)\,\mathrm{d}x.$$

Then the covariance of G_1 and G_2 is

$$\mathrm{cov}\{G_1(B), G_2(B)\} = H(B)\{1 - H(B)\}M_3 \int_0^\infty \int_0^\infty \beta(v_1, v_2; M_1, M_2, M_3)\,\mathrm{d}v_1\,\mathrm{d}v_2,$$

where

$$\beta(v_1, v_2; M_1, M_2, M_3) = -L_\eta''(v_1 + v_2)\exp\{-M_3 L_\eta(v_1 + v_2) - M_1 L_\eta(v_1) - M_2 L_\eta(v_2)\}.$$

For a proof, see Appendix A.

Similarly, expressions can be derived for $\mathrm{var}\{G_1(B)\}$ and $\mathrm{var}\{G_2(B)\}$ and so

$$\rho = \mathrm{corr}(G_1, G_2) = \frac{M_3 \int_0^\infty \int_0^\infty \beta(v_1, v_2; M_1, M_2, M_3)\,\mathrm{d}v_1\,\mathrm{d}v_2}{\sqrt{\{(M_1 + M_3)(M_2 + M_3)\,\beta^*(M_1 + M_3)\,\beta^*(M_2 + M_3)\}}},$$

where

$$\beta^*(M) = \int_0^\infty \int_0^\infty -L_\eta''(v_1 + v_2)\exp\{-M L_\eta(v_1 + v_2)\}\,\mathrm{d}v_1\,\mathrm{d}v_2.$$

In the special case where $M_3 = M\rho^*$ and $M_1 = M_2 = M(1-\rho^*)$ for $0 < \rho^* < 1$, we obtain

$$\rho = \rho^*(1+\varepsilon),$$

where

$$\varepsilon = \frac{\int_0^\infty \int_0^\infty -L''_\eta(v_1+v_2)\exp\{-M\,L_\eta(v_1+v_2)\}\gamma(v_1,v_2)\,\mathrm{d}v_1\,\mathrm{d}v_2}{\beta^*(M)},$$

with

$$\gamma(v_1,v_2) = \exp[-M(1-\rho^*)\{L_\eta(v_1)+L_\eta(v_2)-L_\eta(v_1+v_2\}] - 1.$$

Therefore, the correlation between G_1 and G_2, ρ, can be approximated by ρ^* if $\gamma(v_1,v_2)$ is close to 0 for all ρ^*. This simple result is intuitively appealing since ρ^* reflects the relative importance of the shared component and larger contributions of the shared component will lead to more closely related distributions. It is important to point out that we do not necessarily advocate adopting the restricted parameterization for M_1, M_2 and M_3 in the special case that is used above (with the restriction $M_1 = M_2$), but it is a useful device to understand the properties of our models better. The usefulness of the approximation is illustrated in the following example, using this restricted parameterization.

2.2.1. Example: (correlated) normalized generalized gamma process marginals

In this case $L_\eta(v) = (1/a)\{(v+\lambda)^a - \lambda^a\}$ and $L''_\eta(v) = (a-1)(v+\lambda)^{a-2}$, which implies that

$$\gamma(v_1,v_2) = \exp\left[-M(1-\rho^*)\frac{1}{a}\{(v_1+\lambda)^a + (v_2+\lambda)^a - (v_1+v_2+\lambda)^a - \lambda^a\}\right] - 1.$$

The expression for the Dirichlet process marginals can be found as the limit as $a \downarrow 0$ and using $\lambda = 1$. Fig. 1 shows the relationship between ρ^* and the actual correlation ρ for CNGG processes with different choices of the parameters (including the CDP case in Figs 1(a)–1(c)). The correlation is close to ρ^* for each choice of the hyperparameters with the largest differences for the smaller values of a and λ.

For general M_1, M_2 and M_3, these results suggest that increasing M_3 relative to M_1 and M_2 leads to a larger correlation between G_1 and G_2. When $q > 2$, we can always write a pair of unnormalized distributions $\tilde{G}_j$ and $\tilde{G}_k$, where $j \neq k$, as

$$\tilde{G}_j = \tilde{G}^{(c)} + \tilde{G}^{(j)},$$
$$\tilde{G}_k = \tilde{G}^{(c)} + \tilde{G}^{(k)},$$

where, using $I(\cdot)$ to denote the indicator function, the Lévy measure of $\tilde{G}^{(c)}$ is given by $\{\Sigma_{m=1}^p I(D_{jm}=1, D_{km}=1)M_m\}\eta(x)$, $\tilde{G}^{(j)}$ has Lévy measure $\{\Sigma_{m=1}^p I(D_{jm}=1, D_{km}=0)M_m\}\eta(x)$ and $\tilde{G}^{(k)}$ has Lévy measure $\{\Sigma_{m=1}^p I(D_{jm}=0, D_{km}=1)M_m\}\eta(x)$. This suggests using the general approximation

$$\mathrm{corr}(G_j, G_k) \approx \frac{M^{(c)}}{\sqrt{(M^{(c)}+M^{(j)})}\sqrt{(M^{(c)}+M^{(k)})}}, \tag{5}$$

where $M^{(c)} = \Sigma_{m=1}^p I(D_{jm}=1, D_{km}=1)M_m$, $M^{(j)} = \Sigma_{m=1}^p I(D_{jm}=1, D_{km}=0)M_m$ and $M^{(k)} =$

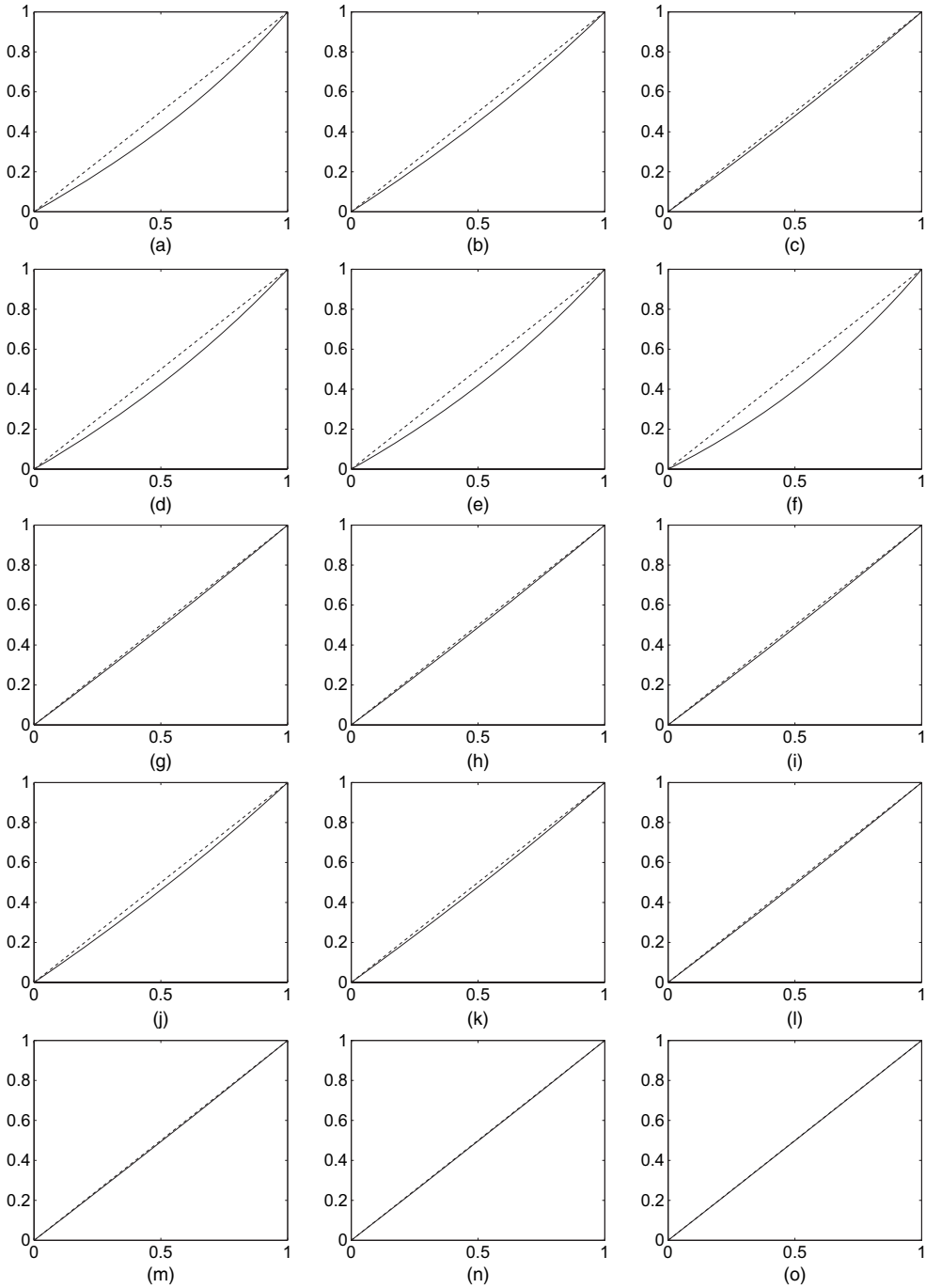

Fig. 1. Plot of the actual correlation ρ (———) and ρ^* (-------) against ρ^* for the CNGG process with various values of a, λ and M: (a) $a = 0$, $\lambda = 1$, $M = 1$; (b) $a = 0$, $\lambda = 1$, $M = 3$; (c) $a = 0$, $\lambda = 1$, $M = 10$; (d) $a = 0.5$, $\lambda = 0, M = 1$; (e) $a = 0.5$, $\lambda = 0, M = 3$; (f) $a = 0.5$, $\lambda = 0, M = 10$; (g) $a = 0.9$, $\lambda = 0, M = 1$; (h) $a = 0.9$, $\lambda = 0, M = 3$; (i) $a = 0.9$, $\lambda = 0, M = 10$; (j) $a = 0.5$, $\lambda = 1, M = 1$; (k) $a = 0.5$, $\lambda = 1, M = 3$; (l) $a = 0.5$, $\lambda = 1, M = 10$; (m) $a = 0.9$, $\lambda = 1, M = 1$; (n) $a = 0.9$, $\lambda = 1, M = 3$; (o) $a = 0.9$, $\lambda = 1, M = 10$

$\Sigma_{m=1}^{p} I(D_{jm}=0, D_{km}=1)M_m$. Therefore, $\mathrm{corr}(G_j, G_k)$ increases as the value of $M^{(c)}$ increases relative to $M^{(j)}$ and $M^{(k)}$. Generally, increasing M_h leads to increased correlations between all distributions with a 1 in the hth column of D.

2.3. Partition probability functions

The previous subsection describes some of the properties of the CNRMI prior. The posterior properties are also interesting and an important step to understanding them is the derivation of the partition probability function which describes the pattern of ties in a sample drawn from distributions with a CNRMI prior. This is known as the exchangeable partition probability function when we have a single sample which can be considered exchangeable. However, the CNRMI prior defines exchangeable sequences in each group but not in the whole sample and so we refrain from using the term exchangeable partition probability function.

Theorem 2. Suppose that H is a non-atomic probability distribution, n_g observations have been taken in the gth group, there are K distinct values across all samples and $n_{j,i}$ is the number of times that the ith distinct value is observed in the sample for the jth group. Let $f^{(i)}$ be a q-dimensional vector for which

$$f_j^{(i)} = \begin{cases} 1 & \text{if } n_{j,i} > 0, \\ 0 & \text{otherwise,} \end{cases}$$

and let a_i be the indices of all columns of D that can be formed by replacing 0s with 1s in $f^{(i)}$ (including $f^{(i)}$ if it is a column of D). The unknown index of the underlying measure that generated the jth distinct value is denoted by $z_j \in \{1, \ldots, p\}$. Define $\mathcal{Z} = \times_{i=1}^{K} a_i$,

$$K_j^*(z) = \#\{i | z_i = j\},$$

$$m_i = \sum_{j=1}^{q} n_{j,i}$$

and

$$I_\eta(n, v) = \int J^n \exp(-vJ)\,\eta(J)\,\mathrm{d}J.$$

The partition probability function, describing the probability of obtaining a particular random partition, is then given by

$$\prod_{g=1}^{q} \frac{1}{\Gamma(n_g)} \sum_{z \in \mathcal{Z}} \int_{(0,\infty)^q} v_g^{n_g - 1} \prod_{j=1}^{p} M_j^{K_j^*(z)} \prod_{i=1}^{K} I_\eta\left(m_i, \sum_{g=1}^{q} D_{gz_i} v_g\right) \prod_{j=1}^{p} \exp\left\{-M_j L_\eta\left(\sum_{g=1}^{q} D_{g_j} v_g\right)\right\} \mathrm{d}v,$$

where $v = (v_1, \ldots, v_q)$.

Lijoi *et al.* (2011) derived the same expression (but with rather different notation) for the case of two groups. In general, the integral in this expression will be difficult to evaluate analytically. However, an analytic expression can be derived in the special case of a CDP prior with two groups.

Corollary 1. If we assume the CDP prior with $q = 2$, $p = 3$ and

$$D = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix},$$

the partition probability function is

$$\sum_{z\in\mathcal{Z}}\prod_{i=1}^{K}\Gamma(m_i)\prod_{j=1}^{3}M_j^{K_j^*(z)}\frac{\Gamma(M_1^*+M_3^*-n_1)}{\Gamma(M_1^*+M_3^*)}\frac{\Gamma(M_2^*+M_3^*-n_2)}{\Gamma(M_2^*+M_3^*)}$$
$$\times {}_3F_2(M_3^*,n_1,n_2;M_1^*+M_3^*,M_2^*+M_3^*;1),$$

where ${}_qF_p$ is the generalized hypergeometric function and $M_i^*=M_i+m_i$.

This result can be interpreted as follows. Using the notation of theorem 1, the distinct values in the first group must be generated from either $\tilde{G}_1^*$ or $\tilde{G}_3^*$ and, similarly, the distinct values in the second group must be generated from either $\tilde{G}_2^*$ or $\tilde{G}_3^*$. If these assignments were known then the partition probability function could be easily calculated. The result in corollary 1 arises from summing this expression over all possible ways that the distinct values could be generated, which are given by the set $\mathcal{Z}$.

In the special cases of the CDP prior and normalized stable marginals with two groups, Lijoi *et al.* (2011) derived expressions for a Pólya urn representation.

2.4. Modelling of groups

In the simple case with two groups, there are naturally three underlying random measures $\tilde{G}_j^*$ in our model: one modelling the common mass shared between the groups and two for the idiosyncratic components. In cases with more groups, we need to make modelling decisions, which are more fully explored in this subsection. The most flexible models in our class are generated by allocating a separate random measure for modelling the mass that is shared by each non-empty subset of group distributions. The most complete model for q groups in the CNRMI class with a given M, H and η can thus be defined by taking $p=2^q-1$ and letting the ith column of D be the binary representation of i for $1\leqslant i\leqslant 2^q-1$. For example, if $q=3$, then

$$D=\begin{pmatrix}0&0&0&1&1&1&1\\0&1&1&0&0&1&1\\1&0&1&0&1&0&1\end{pmatrix}, \tag{6}$$

where $\tilde{G}_1^*$, $\tilde{G}_2^*$ and $\tilde{G}_4^*$ are idiosyncratic components, $\tilde{G}_3^*$, $\tilde{G}_5^*$ and $\tilde{G}_6^*$ are shared by two groups and $\tilde{G}_7^*$ is shared by all three groups. This will be called the saturated model. The levels of correlation between the distributions can be accommodated by choosing appropriate values of $M_1,\ldots,M_p$ and using approximation (5). Clearly this model becomes increasingly complicated as q increases. More parsimonious models can be constructed by removing columns of D from the saturated model (which is equivalent to setting some M_h to 0). A model with a similar structure to that of Müller *et al.* (2004), with a single common component and q idiosyncratic components, would use the $q\times(q+1)$-dimensional matrix

$$D=(\mathbf{1}_q \quad I_q),$$

where $\mathbf{1}_q$ is a q-dimensional vector of 1s (representing the single common component) and I_q is the $(q\times q)$-dimensional identity matrix. Alternatively, if distributions at different times are being modelled then a simple model could be defined using

$$D=(\mathbf{1}_q \quad I_q \quad R),$$

where R is a $q\times(q-1)$-dimensional matrix for which $R_{ij}=1$ if $j=i$ or $j=i-1$ and $R_{ij}=0$ otherwise. The model then includes a common underlying measure (in the first column), idiosyncratic underlying measures (in the next q columns) and underlying random measures shared by consecutive distributions (in the next $q-1$ columns). More problem-specific forms of dependence could also be modelled. Suppose that we take observations from three distributions

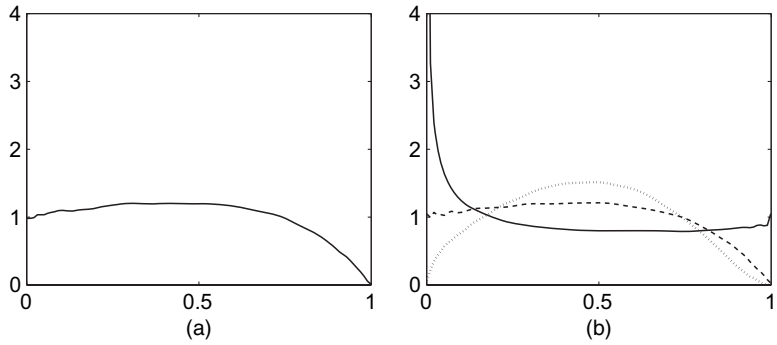

Fig. 2. The prior on $\rho = \text{corr}(G_1, G_2)$ for the saturated model with $q = 2$ and (a) a point mass prior on M_h and (b) a shrinkage prior with $M_h \sim \text{Ga}(M^*/2, \phi)$ where $M^* = 1$ (———), $M^* = 2$ (-------) and $M^* = 3$ (·····) and $\phi = 1$

where we think that G_1 and G_2 are more related to each other than to G_3. A suitable model would have

$$D = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix},$$

where the inclusion of the final column allows extra dependence between G_1 and G_2.

In practice, we may not have prior information that leads us to consider models that are simpler than the saturated model. We suggest using regularization to avoid overfitting (since the number of underlying processes p grows quickly with q). A standard approach would be Bayesian variable selection on the columns of D for the saturated model. Setting $M_h = 0$ in the Lévy measure of the underlying measure $\tilde{G}_h^*$ leads to a degenerate Lévy process with no jumps and so is equivalent to excluding column h. We use this formal model selection strategy with a prior point mass at zero. While always imposing that each group has a properly defined random probability measure (i.e. B_j as defined in Section 2.1 is non-zero for all q groups), we assume that $p(M_h \neq 0) = 2^{1-q}$. This ensures that the prior mass on the correlation between two distributions being 0 does not tend to 0 or 1 as the number of groups increases. If $M_h \neq 0$, then it is chosen from an exponential distribution with mean 1. Under this point mass prior, the probability that the correlation between any two distributions is 0 depends on q and in the saturated case it is equal to 0.2 when $q = 2$ and is approximately 0.2 for larger q. Fig. 2(a) displays the induced prior mass on non-zero values of ρ, the correlation coefficient between G_1 and G_2, when $q = 2$ and D as in corollary 1, which shows a dispersed distribution on $(0, 1)$. In our applications, we shall mostly focus on an alternative approach and define a prior for M_h which encourages substantial shrinkage towards zero (this is similar to the shrinkage prior approach to regression described by Scott and Polson (2011) and Griffin and Brown (2010)). The prior for $M_1, M_2, \ldots, M_p$ is chosen in the following way. The values of $M_1, M_2, \ldots, M_p$ control the dependence between distributions and can be chosen to represent prior beliefs. The additive effect of the M_js is useful here. Suppose that we have one distribution with M chosen to take the value M^*. Moving to two distributions and assuming that G_1 and G_2 have the same distribution, the saturated model suggests that $M_1 + M_3 = M^*$ and $M_2 + M_3 = M^*$, so $M_1 = M_2$. If we are indifferent between an observation being allocated to a shared cluster or an idiosyncratic cluster then $M_1 = M_2 = M_3$. Repeated use of this argument allows extension to any value of q and suggests that $M_1, M_2, \ldots, M_p$ are independent and $M_i \sim \text{Ga}(M^*/2^{q-1}, \phi)$. Fig. 2(b) shows the prior density function induced for the correlation ρ in the saturated case

with $q=2$ for various values of M^*. All priors are centred on $\frac{1}{2}$ with the variability decreasing as M^* increases. We shall mainly focus on $M^*=1$ in our applications. This prior is relatively flat for correlations larger than 0.1 and has larger mass close to zero. This will lead to some shrinkage of small correlations.

3. Computational methods

This section describes an MCMC sampler for fitting the general mixture model

$$y_{g,i} \sim k(y_{g,i}|\theta_{g,i}), \qquad i=1,2,\ldots,n_g,$$

$$\theta_{g,i} \sim G_g \qquad \text{for group } g=1,\ldots,q$$

and

$$G_1, G_2, \ldots, G_q \sim \text{CNRMI}(M, H, D; \eta),$$

where H and η potentially have hyperparameters which also have priors and H has density h. The total number of observations is $n=\Sigma_{g=1}^{q} n_g$. First, we describe the algorithm for the prior $M_i \sim \text{Ga}(\tau_i, \phi_i)$. The changes to the sampler needed for the prior with $p(M_i=0)=p_0$ and $M_i \sim \text{Ga}(\tau_i, \phi_i)$ with probability $1-p_0$ are described in Section 3.8.

Walker (2007) introduced a slice sampling method to sample from the posterior of a Dirichlet process mixture model without truncating its stick breaking representation. Kalli *et al.* (2011) developed a more efficient version of this algorithm and extended it to several non-parametric priors. Several slice sampling algorithms for normalized random-measure mixture models were introduced by Griffin and Walker (2011). They showed that their algorithms are competitive with other algorithms for non-conjugate non-parametric mixture models. We shall extend their 'Slice 1' algorithm. Importantly, all full conditional distributions involve only a finite number of jumps from the infinite activity Lévy processes but avoid directly truncating smaller jumps of the Lévy process, unlike other methods for non-conjugate NRMI mixtures (Nieto-Barajas and Prünster, 2009). It should be noted that simply normalizing the jumps of a compound Poisson process does not generate a well-defined non-parametric process. For a single normalized random-measure mixture the posterior is proportional to

$$p(J) \prod_{j=1}^{\infty} h(\theta_j) \prod_{i=1}^{n} w_{s_i} k(y_i|\theta_{s_i}),$$

where $w_i = J_i / \Sigma_{l=1}^{\infty} J_l$, $J=(J_1, J_2, J_3, \ldots)$ and $\theta=(\theta_1, \theta_2, \theta_3, \ldots)$. Griffin and Walker (2011) demonstrated that the following posterior with additional auxiliary variables $u_1, u_2, \ldots, u_n$ and $v_1, v_2, \ldots, v_n$ and integrating over all jumps smaller than $L=\min\{u_i\}$ is a much simpler form for computational purposes:

$$p(J^{(L)}) \prod_{j=1}^{K} h(\theta_j^{(L)}) \prod_{i=1}^{n} I(u_i < J_{s_i}^{(L)}) \exp\left(-v_i \sum_{l=1}^{K} J_l^{(L)}\right) E\left[\exp\left(-v_i \sum_{l=1}^{\infty} J_l^{(-L)}\right)\right] k(y_i|\theta_{s_i}^{(L)}),$$

where $J^{(L)}=\{J_l | J_l \geqslant L\}$, which has K elements, $\theta^{(L)}$ is a vector of associated locations, $J^{(-L)}=\{J_l | J_l < L\}$, $u_1, u_2, \ldots, u_n > 0$ and $v_1, v_2, \ldots, v_n > 0$. The expectation can be evaluated by using the Lévy–Khintchine formula and so

$$E\left[\exp\left(-v \sum_{i=1}^{\infty} J_i^{(-L)}\right)\right] = \exp\left[-M \int_0^L \{1-\exp(-vx)\}\, \eta(x)\, \mathrm{d}x\right].$$

The integral in the exponential is sometimes available in terms of special functions (this is the case for the Dirichlet process) or can be evaluated by using standard quadrature methods.

A latent representation of the posterior of a mixture model using the weights in Section 2 can be derived in a suitable form for computation by introducing latent variables $\{s_{j,i}\}_{j=1:q,i=1:n_j}$ which are allocation variables for mixture components whereas $\{r_{j,i}\}_{j=1:q,i=1:n_j}$ allocates each observation to one of p underlying random measures $\tilde{G}_1^*, \tilde{G}_2^*, \ldots, \tilde{G}_p^*$. Thus, the observation $y_{g,i}$ is assumed to be drawn from $k(\cdot|\theta^{(L)}_{r_{g,i},s_{g,i}})$. Using auxiliary variables $u_{j,1}, \ldots, u_{j,n_j}$ and $v_{j,1}, \ldots, v_{j,n_j}$ for group j, the posterior can now be expressed as

$$\prod_{j=1}^{q} \frac{V_j^{n_j-1}}{\Gamma(n_j)} \prod_{i=1}^{p} p(J_i^{(L)}) \prod_{i=1}^{p} \prod_{l=1}^{K_i} h(\theta_{i,l}^{(L)}) \prod_{j=1}^{q} \prod_{i=1}^{n_j} I(u_{j,i} < J^{(L)}_{r_{j,i},s_{j,i}}) k(y_{j,i}|\theta^{(L)}_{r_{j,i},s_{j,i}}) \times \exp(-V^{\mathrm{T}} DJ^{(+)}) E[\exp(-V^{\mathrm{T}} DJ^{(\infty)})], \quad (7)$$

where $J_i^{(L)} = \{J_{i,l} | J_{i,l} \geqslant L\}$, which has K_i elements, and $\theta_i^{(L)}$ is a vector of associated locations. We also define V to be a q-dimensional vector where $V_i = \Sigma_{j=1}^{n_i} v_{i,j}$, $J^{(+)}$ is a p-dimensional vector with $J_i^{(+)} = \Sigma_{l=1}^{K_i} J_{i,l}^{(L)}$ and $J^{(\infty)}$ is a p-dimensional vector where $J_i^{(\infty)} = \Sigma_{l=1}^{\infty} J_{i,l} - \Sigma_{l=1}^{K_i} J_{i,l}^{(L)}$. We can show that $K_i \sim \mathrm{Pn}\{M_i \int_L^{\infty} \eta(x)\,\mathrm{d}x\}$ ($\mathrm{Pn}(b)$ denotes a Poisson distribution with mean b) and $J_1, J_2, \ldots, J_{K_i}$ are independent conditional on K_i with $p(J_{i,l}^{(L)}) = \eta(J_{i,l}^{(L)}) / \int_L^{\infty} \eta(x)\,\mathrm{d}x$ where $J_{i,l}^{(L)} \in (L, \infty)$ (so that $J_i^{(L)}$ follows a compound Poisson process). Integrating out $u_{j,i}$ in expression (7), we obtain

$$\prod_{j=1}^{q} \frac{V_j^{n_j-1}}{\Gamma(n_j)} \prod_{i=1}^{p} p(J_i^{(L)}) \prod_{j=1}^{p} \prod_{i=1}^{K_j} h(\theta_{j,i}^{(L)}) J_{j,i}^{(L)m_{j,i}} \prod_{j=1}^{q} \prod_{i=1}^{n_j} k(y_{j,i}|\theta^{(L)}_{r_{j,i},s_{j,i}}) \times \exp(-V^{\mathrm{T}} DJ^{(+)}) E[\exp(-V^{\mathrm{T}} DJ^{(\infty)})],$$

where $m_{j,i} = \#\{(l,k) | s_{l,k} = i \text{ and } r_{l,k} = j, 1 \leqslant k \leqslant n_l, 1 \leqslant l \leqslant q\}$ is the size of the cluster of observations that are associated with $\theta_{j,i}^{(L)}$.

Each expectation in the product can be evaluated by using the Lévy–Khintchine formula, so

$$E[\exp(-V^{\mathrm{T}} DJ^{(\infty)})] = \exp(-\mathbf{1}_p^{\mathrm{T}} \tilde{M} \tilde{E}),$$

where $\tilde{M}$ is a $p \times p$ diagonal matrix with $\tilde{M}_{hh} = M_h$ and, defining $D_{\cdot i}$ as the ith column of D, $\tilde{E}$ is the p-dimensional vector with ith element

$$\tilde{E}_i = \int_0^L \{1 - \exp(-V^{\mathrm{T}} D_{\cdot i} x)\} \eta(x)\,\mathrm{d}x.$$

Therefore the latent representation of the posterior retains much of the linearity that is introduced in the model. The chain can be initialized in the following way. Choose a starting truncation point L and generate $K_i \sim \mathrm{Pn}\{M_i \int_L^{\infty} \eta(x)\,\mathrm{d}x\}$; then $J_i^{(L)}$ are simulated from the distribution with density $p(J_{i,l}^{(L)}) = \eta(J_{i,l}^{(L)}) / \int_L^{\infty} \eta(x)\,\mathrm{d}x$ where $J_{i,l}^{(L)} \in (L, \infty)$. The locations $\theta_{j,1}^{(L)}, \theta_{j,2}^{(L)}, \ldots, \theta_{j,K_j}^{(L)}$ are taken to be independent and identically distributed from H and the latent variables $r_{j,i}$ and $s_{j,i}$ can be simulated from the discrete distributions

$$p(r_{j,i} = k) = D_{jk} M_k \Big/ \sum_{l=1}^{p} D_{jl} M_l, \qquad k = 1, 2, \ldots, p,$$

and

$$p(s_{j,i} = k) \propto k(y_i|\theta^{(L)}_{r_{j,i},k}) J^{(L)}_{r_{j,i},k}, \qquad 1 \leqslant k \leqslant K_{r_{j,i}}.$$

The slice latent variables $u_{j,i}$ are uniformly distributed on $(L, J^{(L)}_{r_{j,i},s_{j,i}})$.

In the following steps, we define $J^*_j = \{J_{j,i} | m_{j,i} \neq 0\}$, i.e. the jumps in the jth component process which have observations allocated to them. Recall that $m_{j,i}$ is the number of observations allocated to jump $J_{j,i}$. The steps of the Gibbs sampler are described for the case where all M_h have a gamma distribution and are as follows.

3.1. Step 1: split–merge move

The problem of multimodality of the posterior distribution in these models and a computational solution, a split–merge move, are described in Kolossiatis *et al.* (2012). In our model, it is useful to link the underlying measures to their corresponding columns in the D-matrix. For example, in the saturated model with $q = 3$ and D given in equation (6), the underlying random measure $\tilde{G}^*_1$ will be referred to as the 'underlying random measure $(0, 0, 1)$'. The split–merge move is performed in the following way. A split move is selected with probability $\frac{1}{2}$; otherwise a merge move is proposed. An underlying random measure $\mathbf{e}$, a column of D, is selected at random from those underlying random measures which have observation allocated to them and a non-empty mixture component i^* from $\mathbf{e}$ is selected uniformly at random. If the split move is selected, the members of the cluster are divided according to their group membership into two clusters $\mathbf{e}_1$ and $\mathbf{e}_2$. For example, in the saturated model with $q = 3$, if we choose $\mathbf{e} = (1, 1, 0)$ the cluster would be split into a cluster in the underlying measure $(1, 0, 0)$ and a cluster in the underlying measure $(0, 1, 0)$. In this case, there is only one possible split. However, if we choose $\mathbf{e} = (1, 1, 1)$, there are three possible splits: clusters in $(1, 0, 0)$ and $(0, 1, 1)$, clusters in $(0, 1, 0)$ and $(1, 0, 1)$ or clusters in $(0, 0, 1)$ and $(1, 1, 0)$. The particular split is chosen uniformly at random from all possible splits. The merge move performs the opposite operation. For this move, a set of allowable underlying measures is defined, $\mathcal{C} = \{\mathbf{e}^* | \mathbf{e}^*_j = 0 \text{ for all } j \text{ for which } \mathbf{e}_j = 1\}$, and an underlying measure $\mathbf{e}^\dagger$ is chosen uniformly at random from $\mathcal{C}$ (this happens regardless of whether there are any non-empty clusters allocated to that measure). One of the non-empty clusters, j^*, in $\mathbf{e}^\dagger$ (if any exist) or a 'null' cluster is chosen at random with equal probability. A new cluster is then formed in the underlying random measure $\mathbf{e}^{\text{comb}}$ where $\mathbf{e}^{\text{comb}}_i = 1$ if $\mathbf{e}_i = 1$ or $\mathbf{e}^\dagger_i = 1$ and $\mathbf{e}^{\text{comb}}_i = 0$ if $\mathbf{e}_i = 0$ and $\mathbf{e}^\dagger_i = 0$. If a null cluster were selected then the cluster i^* is moved from the underlying measure $\mathbf{e}$ to $\mathbf{e}^{\text{comb}}$. Otherwise, clusters i^* and j^* are combined to define a new cluster in $\mathbf{e}^{\text{comb}}$. Let (s, r) and (s', r') be the values of the latent allocation variables before and after making the move respectively. We assume that $M_j \sim \text{Ga}(\tau_j, \phi_j)$. The acceptance probability is calculated by integrating out the jumps and for the split move has the form

$$\max\left\{1, \frac{p(y|s', r')\, p(s', r')}{p(y|s, r)\, p(s, r)} \frac{\kappa\, K^*(\mathbf{e})\, S(\mathbf{e})}{2\kappa'\, K^{*\prime}(\mathbf{e}_1)\, M(\mathbf{e}_1)\{K^{*\prime}(\mathbf{e}_2) + 1\}}\right\},$$

where κ and κ' are the number of underlying random measures with observations allocated to them before and after the move respectively, $K^*(\mathbf{e})$ and $K^{*\prime}(\mathbf{e})$ are the number of non-empty clusters in the random measure $\mathbf{e}$ before and after the move, $M(\mathbf{e})$ is the size of $\mathcal{C}$ and $S(\mathbf{e})$ is the number of pairs of underlying measures that can be formed by splitting $\mathbf{e}$. In addition, we can write

$$p(s, r) = \prod_{j=1}^{p} \frac{\Gamma(\tau_j + K^*_j)}{(\phi_j + \tilde{A}_j)^{\tau_j + K^*_j}} \prod_{\{i | m_{j,i} \neq 0\}} \int J^{m_{j,i}}_{j,i} \exp(-J_{j,i} V^{\mathrm{T}} D_{\cdot i})\, \eta(J_{j,i})\, \mathrm{d}J_{j,i},$$

where for $j = 1, 2, \ldots, p$ we have $K^*_j = \#\{k : m_{j,k} > 0\}$ and

$$\tilde{A}_j = \int_0^\infty \{1-\exp(-V^{\mathrm{T}} D_{\cdot j} x)\}\,\eta(x)\,\mathrm{d}x.$$

The move is completed by sampling M from its full conditional distribution and then sampling u, K and J.

3.2. Step 2: updating V

Defining $\tilde{A} = (\tilde{A}_1, \ldots, \tilde{A}_p)^{\mathrm{T}}$, the full conditional distribution of V_j is proportional to

$$V_j^{n_j - 1} \prod_{l=1}^{K_j} \int J_{j,l}^{m_{j,l}} \exp(-J_{j,l} V^{\mathrm{T}} D_{\cdot j})\,\mathrm{d}J_{j,l} \exp(-\mathbf{1}_p^{\mathrm{T}} \tilde{M} \tilde{A}), \qquad V_j > 0.$$

The parameter can be updated by using a Metropolis–Hastings random walk on the log-scale. We also found it useful to update $V^* = \Sigma_{j=1}^q V_j$ conditionally on $Q = (V_1/V^*, \ldots, V_q/V^*)$. The full conditional distribution of $V^* > 0$ is

$$V^{*n-1} \prod_{j=1}^{p} \prod_{l=1}^{K_j} \int J_{j,l}^{m_{j,l}} \exp(-J_{j,l} V^* Q^{\mathrm{T}} D_{\cdot j})\,\mathrm{d}J_{j,l} \exp(-\mathbf{1}_p^{\mathrm{T}} \tilde{M} \tilde{A}).$$

The parameter can be updated by using a Metropolis–Hastings random walk on the log-scale. If $V^{*\prime}$ is accepted then each V_j is updated to $V_j V^{*\prime}/V^*$.

3.3. Step 3: updating M

The full conditional distribution of M_j is proportional to

$$p(M_j) M_j^{K_j} \exp\left[-M_j \int_0^\infty \{1-\exp(-V_j J)\}\,\eta(J)\,\mathrm{d}J\right]$$

and if $p(M_j) \sim \mathrm{Ga}(\tau_j, \phi_j)$ then the full conditional distribution is

$$\mathrm{Ga}\left[\tau_j + K_j, \phi_j + \int_0^\infty \{1-\exp(-V_j J)\}\,\eta(J)\,\mathrm{d}J\right].$$

3.4. Step 4: updating u and $J_1^{(L)}, J_2^{(L)}, \ldots, J_p^{(L)}$

This set of full conditional distributions can be updated by using the efficient slice sampling method of Kalli *et al.* (2011) by integrating out $u = \{u_{j,i}\}_{j=1:q,\, i=1:n_j}$ when updating the jumps. The update is described for NRMI mixtures by Griffin and Walker (2011) and can be simply extended to our model. The elements of $J_1^*, J_2^*, \ldots, J_p^*$ are simulated first followed by the elements of u (which only depends on J_k through the elements of J_k^*) and finally the other elements of $J_k^{(L)}$. The full conditional distribution of the element $J_{k,l}^{(L)} \in J_k^*$ is proportional to

$$J_{k,l}^{(L) m_{k,l}} \exp(-J_{k,l}^{(L)} V^{\mathrm{T}} D_{\cdot k})\,\eta(J_{k,l}), \qquad J_{k,l}^{(L)} > 0.$$

The full conditional of $u_{j,i}$ is uniform on $(0, J_{r_{j,i}, s_{j,i}}^{(L)})$ and this allows us to calculate $L = \min\{u_{j,i}\}$. Finally, the elements of $J_k^{(L)}$ for which $m_{k,l} = 0$ can be simulated as follows. First, simulate $K_k \sim \mathrm{Pn}\{M_k \int_L^\infty \exp(-V^{\mathrm{T}} D_{\cdot k} x)\,\eta(x)\,\mathrm{d}x\}$; then draw K_k jumps with density proportional to $\eta(J_{k,l}^{(L)}) / \int_L^\infty \eta(x)\,\mathrm{d}x$ and associate a $\theta_{k,l}^{(L)}$ drawn from H with each jump. Some details of simulating the jumps are given in Griffin and Walker (2011).

3.5. Step 5: updating $\theta^{(L)}$

The elements of $\theta^{(L)}$ are independent under their joint full conditional distribution, and the density of $\theta_{l,k}^{(L)}$ is proportional to

$$h(\theta_{l,k}^{(L)}) \prod_{\{(j,i)|s_{j,i}=k \text{ and } r_{j,i}=l\}} k(y_{j,i}|\theta_{l,k}^{(L)}),$$

which is a familiar form that is used in samplers for many infinite mixture models, such as Dirichlet process mixtures.

3.6. Step 6: updating s and r

The latent variables $s_{j,i}$ and $r_{j,i}$ can be updated jointly and drawn from their full conditional distribution

$$p(s_{j,i}=k \text{ and } r_{j,i}=l) \propto D_{jl}\, I(J_{l,k}^{(L)} > u_{j,i})\, k(y_{j,i}|\theta_{l,k}^{(L)}),$$

where $\{(l,k) : J_{l,k}^{(L)} > u_{j,i}\}$ is a finite set.

3.7. Example: (correlated) normalized generalized gamma process marginals

The Lévy density has the form

$$\eta(x) = \frac{1}{\Gamma(1-a)} x^{-1-a} \exp(-\lambda x).$$

The quantities that are used by the MCMC sampler are

$$\int J_{j,i}^{m_{j,i}} \exp(-J_{j,i} V^{\mathrm{T}} D_{\cdot j})\, \eta(J_{j,i})\, \mathrm{d}J_{j,i} = \frac{1}{\Gamma(1-a)} \frac{\Gamma(m_{j,i}-a)}{(\lambda + V^{\mathrm{T}} D_{\cdot j})^{m_{j,i}-a}},$$

and the full conditional distribution of $J_{j,i}$ is $\mathrm{Ga}(m_{j,i}+a, \lambda + V^{\mathrm{T}} D_{\cdot j})$.

3.8. Markov chain Monte Carlo sampling with a prior point mass at zero for M_h

The model when M_h is given a prior with a point mass at zero can be fitted by using an MCMC sampler which is similar to the one above. A latent variable $\chi = (\chi_1, \chi_2, \ldots, \chi_p)$ is introduced to indicate whether $M_j = 0$ (when $\chi_j = 0$) or $M_j \sim \mathrm{Ga}(\tau_j, \phi_j)$ is non-zero (when $\chi_j = 1$). The differences are described below.

3.8.1. Step 1a: split–merge move

The split–merge move now also includes updating of the latent variable χ. In the split move, the χs that are associated with the underlying random measures in which the new clusters are formed are proposed to be both 1. If, following the split move, there are no clusters in the underlying random measure where the split occurs then χ is set to 0 or 1 with probability $\frac{1}{2}$ each. Otherwise χ is set to 1. Similarly, in the merge move, the underlying random measure to which the new cluster is given has a proposed χ_j equal to 1. The underlying random measures from which the clusters to be merged are drawn can have their χs set to 0 if there are no clusters in that underlying random measure after that move. This again occurs with probability $\frac{1}{2}$. The acceptance rate can be calculated by using

$$p(s,r)=\prod_{\{j|\chi_j=1\}}\frac{\Gamma(\tau_j+K_j^*)}{(\phi_j+\tilde{A}_j)^{\tau_j+K_j^*}}\frac{\phi_j^{\tau_j}}{\Gamma(\tau_j)}\prod_{\{i|m_{j,i}\neq 0\}}\int J_{j,i}^{m_{j,i}}\exp(-J_{j,i}V^{\mathrm{T}}D_{\cdot i})\eta(J_{j,i})\,\mathrm{d}J_{j,i}.$$

3.8.2. Step 3a: updating M

The full conditional of M_j is $\mathrm{Ga}[\tau_j+K_j,\phi_j+\int_0^\infty\{1-\exp(-V_jJ)\}\eta(J)\,\mathrm{d}J]$ if $\chi_j=1$. Otherwise $M_j=0$. In the applications we adopt $\tau_j=\phi_j=1$ throughout.

4. Comparing distributions

Once we have a posterior distribution on the distributions $G_1,G_2,\ldots,G_q$, it is useful to have some graphical summaries which help us to understand the differences between distributions. Most simply, we can write

$$G_i=\bar{G}+\Pi_i,$$

where

$$\bar{G}=\frac{1}{q}\sum_{j=1}^{q}G_j$$

is a 'grand mean' distribution and $\Pi_i=G_i-\bar{G}$ is a signed measure which gives measure zero to Ω and which represents the difference of each distribution from the grand mean. This idea is similar to the modelling of continuous responses in a one-way ANOVA model. Analogies to higher order ANOVA models are also possible. Suppose that the groups are defined by two covariates (x_1 and x_2) and the distribution for the ith level of x_1 ($i=1,\ldots,n$) and the jth level of x_2 ($j=1,\ldots,m$) is represented as $G_{i,j}$. Then we can decompose

$$G_{i,j}=\bar{G}+\Pi_{i\cdot}+\Pi_{\cdot j}+\Gamma_{i,j}, \tag{8}$$

where

$$\bar{G}=\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}G_{i,j},$$

$$\Pi_{i\cdot}=\frac{1}{m}\sum_{j=1}^{m}(G_{i,j}-\bar{G}),$$

$$\Pi_{\cdot j}=\frac{1}{n}\sum_{i=1}^{n}(G_{i,j}-\bar{G})$$

and $\Gamma_{i,j}=G_{i,j}-\bar{G}-\Pi_{i\cdot}-\Pi_{\cdot j}$. Here $\bar{G}$ is a probability measure and $\Pi_{i\cdot}$, $\Pi_{\cdot j}$ and $\Gamma_{i,j}$ are signed measures that put measure 0 on Ω. This separates the effect of level i of x_1 averaged over all levels of x_2, denoted by $\Pi_{i\cdot}$, the average effect of level j of x_2 ($\Pi_{\cdot j}$) and the interaction effects of combinations of levels i and j of both variables ($\Gamma_{i,j}$), giving us a very useful decomposition of the differences between the distributions.

The summaries described so far allow us to understand and interpret the differences between distributions but we also want to say something meaningful about regions of the support where the distributions are particularly different. We shall consider a pair of distributions G_i and G_j, and find a partition $\mathcal{P}$ of Ω defining subsets $\mathcal{P}_k$ and an indicator vector d for which $d_k=-1$ if G_i places substantially more mass than G_j on $\mathcal{P}_k$, $d_k=1$ if G_j places substantially more

mass than G_i on $\mathcal{P}_k$ and $d_k = 0$ otherwise. The choice of $\mathcal{P}$ and d will be made by specifying a utility function and finding the partition that maximizes expected utility. The utility function is

$$U(\mathcal{P}, d) = \sum_{k=1}^{r} U^*(\mathcal{P}_k, d_k),$$

where $\mathcal{P}_1, \ldots, \mathcal{P}_r$ are the elements of $\mathcal{P}$ and

$$U^*(\mathcal{P}, d) = \begin{cases} G_i(\mathcal{P}) - G_j(\mathcal{P}), & d = -1, \\ \frac{\varepsilon}{2}\{G_i(\mathcal{P}) + G_j(\mathcal{P})\}, & d = 0, \\ G_j(\mathcal{P}) - G_i(\mathcal{P}), & d = 1, \end{cases}$$

where $0 < \varepsilon < 2$ is chosen to determine the meaning of substantial difference. Increasing values of ε lead to a utility function that increasingly favours setting $d_k = 0$. To understand the choice of utility function, consider an element $\mathcal{P}_k$ of a fixed partition $\mathcal{P}$. Then, $d_k = 0$ if

$$\frac{|G_i(\mathcal{P}_k) - G_j(\mathcal{P}_k)|}{\frac{1}{2}\{G_i(\mathcal{P}_k) + G_j(\mathcal{P}_k)\}} < \varepsilon.$$

The left-hand side of the expression is the difference in the mass of the two distributions on $\mathcal{P}_k$ divided by the average mass and ε is then interpreted as a tolerance parameter which controls the size of that ratio which constitutes a substantial difference. The expression naturally scales the difference by the mean mass under the two distributions and larger absolute differences will be declared 'similar' in areas with larger average mass.

As $U(\mathcal{P}, d)$ is additive over the elements in the partition, maximizing the utility over partitions is easily done by starting from a very fine partition $\tilde{\mathcal{P}}$ and maximizing U^* on each element. Then we simply join the elements of $\tilde{\mathcal{P}}$ to form the partition $\mathcal{P}$ that maximizes utility.

5. Applications

The methods that are developed in this paper are illustrated on simulated data, a survival analysis example and an example from efficiency measurement. In all cases, the model with NGG marginals with $\lambda = 1$ and unknown other hyperparameters was used. In practice, this is not a particularly restrictive choice. Writing $M = \breve{M}/\lambda^a$ in equation (3) leads to a process where λ scales the jump sizes and so has no effect on the normalized process (we have also implemented inference with a prior on λ and indeed found that the posterior and prior were virtually identical). In all applications, we choose the matrix D corresponding to the saturated model with $p = 2^q - 1$ (even for the stochastic frontier example in Section 5.3, where $q = 6$, so $p = 63$). Throughout, the prior for a was a uniform distribution on $(0, 1)$ and the prior for M_h was $\mathrm{Ga}(M^*/2^{q-1}, 1)$ which implies that the prior for each G_g is NGG with $M \sim \mathrm{Ga}(M^*, 1)$. We shall present results for $M^* = 1$ but also comment on the sensitivity with respect to this choice. Finally, we also discuss results with a prior point mass at zero for M_h.

The MCMC sampler was run with a burn-in of 10000 iterations and a total run of 70000 iterations for all cases. The burn-in period seemed sufficient for the chain to have converged. The split–merge move had an acceptance rate of between 0.1 and 0.2 across the three applications.

5.1. Simulated data

We use two examples to illustrate the flexibility of the model. Example 1 has two groups which each contain 50 data points. The data for the first group are generated from the mixture distribution

$$f_1(x) = \alpha_1 N(x|0,1) + (1-\alpha_1) N(x|-5,1)$$

and in the second group from

$$f_2(x) = \alpha_2 N(x|0,1) + (1-\alpha_2) \mathrm{SkCau}(x|2,2,0.5),$$

where $\mathrm{SkCau}(\cdot|\mu,\sigma,\gamma)$ denotes the skewed Cauchy density function with inverse scale factors as in Fernández and Steel (1998), given by

$$\mathrm{SkCau}(x|\mu,\sigma,\gamma) = \frac{2}{\gamma + 1/\gamma}\{\mathrm{Cau}(x/\gamma|\mu,\sigma)\, I(x>\mu) + \mathrm{Cau}(x\gamma|\mu,\sigma)\, I(x<\mu)\}$$

using $\mathrm{Cau}(\cdot|\mu,\sigma)$ for the standard Cauchy density with location μ and scale σ, and $\gamma > 0$ is the skewness parameter. We adopt $\gamma = 0.5$, which generates considerable negative skewness, whereas the mode remains at $\mu = 2$.

The model of Müller *et al.* (2004) can represent these distributions if $\alpha_1 = \alpha_2$ but that model will fit worse as the values of α_1 and α_2 become further apart. We first consider the case with $\alpha_1 = \alpha_2 = 0.5$.

Example 2 extends the first by defining a third group (so $q = 3$ and D has seven columns) with observations drawn from the same distribution as the second group, so that $f_3(x) = f_2(x)$. In this case, we use $\alpha_1 = 0.5$ and $\alpha_2 = 0.9$. Each data set was fitted by using the model with NGG marginals, given by

$$y_{g,j} \overset{\text{ind}}{\sim} N(\mu_{g,j}, \sigma^2_{g,j}),$$

$$(\mu_{g,j}, \sigma^{-2}_{g,j}) \overset{\text{ind}}{\sim} G_g,$$

$$G_1, G_2, \ldots, G_q \sim \mathrm{CNGG}(M, H, D; a, 1),$$

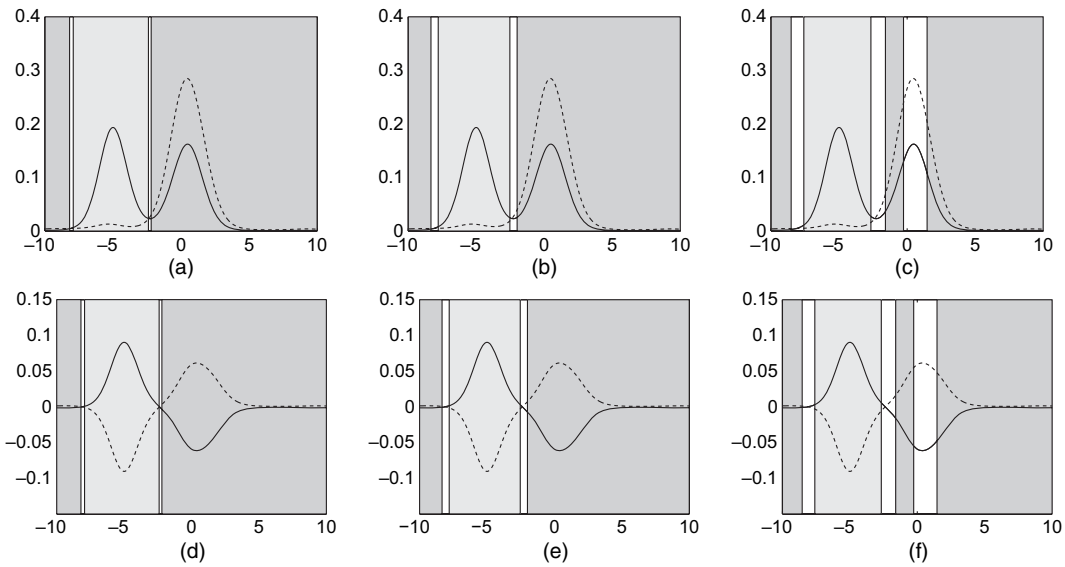

Fig. 3. Example 1 ($\alpha_1 = \alpha_2 = 0.5$)—(a)–(c) posterior predictive densities for the two groups (———, group 1; - - - - - - -, group 2) and (d)–(f) $f_1 - \bar{f}$ (———) and $f_2 - \bar{f}$ (- - - - - - -), indicating the area where group 1 has substantially more mass than group 2 (□) and vice versa (□): (a), (d) $\varepsilon = 0.2$; (b), (e) $\varepsilon = 0.4$; (c), (f) $\varepsilon = 0.6$

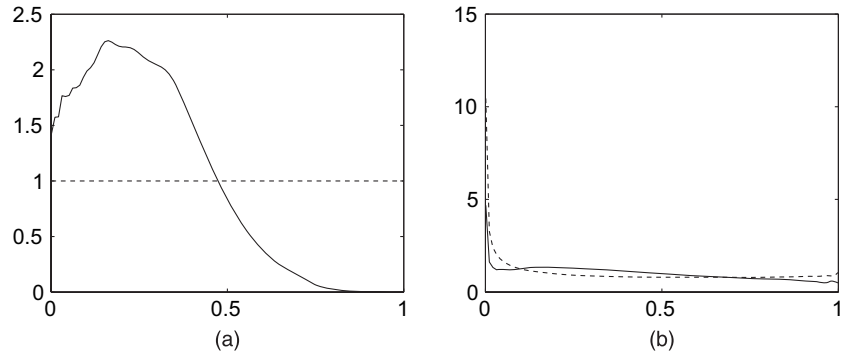

Fig. 4. Example 1 ($\alpha_1 = \alpha_2 = 0.5$): prior (- - - - - - -) and posterior (———) densities of (a) the parameter a and (b) the correlation ρ for the NGG prior

with $H = N(\mu|0, \sigma^2/m_0)\,\mathrm{Ga}(\sigma^{-2}|1, 1)$ where $m_0 = 0.01$.

Some results of fitting the model to data in the first example are shown in Fig. 3. The model estimates the densities well as shown in Figs 3(a)–3(c). The graphs also show partitions of the support found by using the approach in Section 4 for several values of the sensitivity parameter ε. The results are reasonably robust to the choice of ε with $\varepsilon > 0.2$ and they indicate that the distributions are substantially different over most of the region shown. Note that the fat tail and left skewness of the distribution of group 2 lead to the latter having relatively more mass for small x. For $\varepsilon = 0.6$ we no longer distinguish between the mass that is assigned to the two groups in a region around $x = 0$, which, in our view, indicates that this value of ε is perhaps a little large. Figs 3(d)–3(f) show the posterior of the differences between the predictive densities and the overall mean $f_i - \bar{f}$ where $\bar{f} = (1/q)\Sigma_{g=1}^{q} f_g$. It is clear from the definition that $f_1 - \bar{f} = -(f_2 - \bar{f})$ when we have two groups and this is illustrated in the graphs which clearly show where the absolute differences of the densities for the two groups are large.

Fig. 4 shows the posterior densities of the parameter a and the correlation $\rho = \mathrm{corr}(G_1, G_2)$ for the NGG prior. The data favour values of a that are smaller than 0.5 and indicate a preference for general NGG marginal processes over the special cases of Dirichlet process or normalized inverse Gaussian marginals. The posterior distribution of ρ (calculated by using the result of theorem 1) is not very different from the prior, suggesting that the information in the data about correlation is not strong. The mass close to zero is in line with the fact that the distributions that generated both groups are quite different.

The presented findings correspond to $M^* = 1$. Results with $M^* = 2$ and $M^* = 3$ are very close in terms of the predictive distributions and show only minor differences for the posterior on a and ρ. In particular, the posterior mass for a shifts a little towards zero. Posterior results on M_h (which are not reported) are only moderately changed by these changes in the prior and posterior medians of M_h roughly double as M^* changes from 1 to 3.

If we assign priors with point masses at zero to the M_h, we can conduct formal model selection, as mentioned in Section 2.4. With this prior, predictive distributions are virtually identical to those reported in Fig. 3 and the posterior for a is close as well. Posterior inclusion probabilities of the components are close to 1 for the shared component, around 0.8 for the idiosyncratic component of the first group and 0.6 for that of group 2.

Fig. 5 shows results of fitting the model to the second example with three groups. The density estimates clearly show the similarities between groups 2 and 3 and the differences with respect to group 1. The plots of $f_i - \bar{f}$ in Fig. 5(b) clearly illustrate the main differences. The posterior distribution of a is very similar to that shown in Fig. 4.

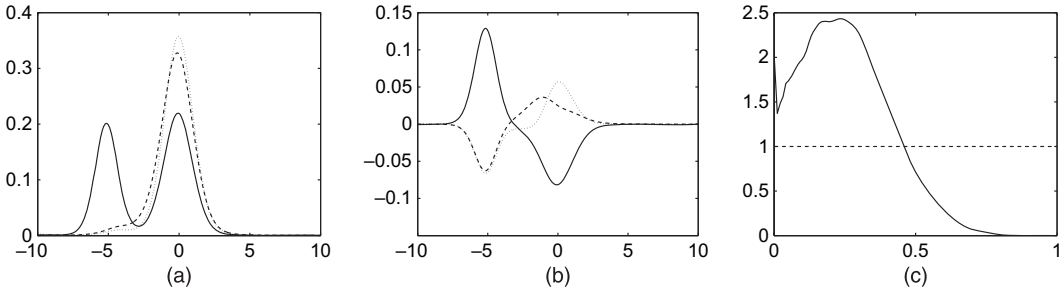

Fig. 5. Example 2 ($\alpha_1 = 0.5; \alpha_2 = 0.9$): (a) posterior predictive densities for the three groups; (b) differences between the predictive densities and the overall mean (———, group 1, $f_1 - \bar{f}$; - - - - - - -, group 2, $f_2 - \bar{f}$; · · · · · · ·, group 3, $f_3 - \bar{f}$); (c) posterior density of a (- - - - - - -, prior)

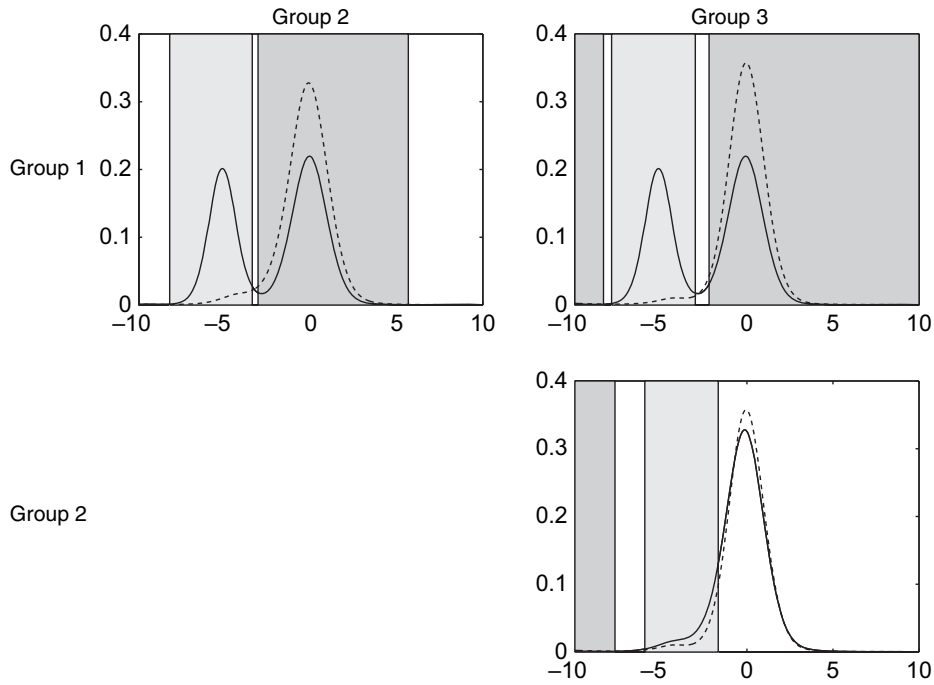

Fig. 6. Example 2 ($\alpha_1 = 0.5; \alpha_2 = 0.9$): posterior mean density for the group in the row (———) and column (- - - - - - -) and comparison of the distributions with dark and light grey areas indicating more mass for respectively the group in the column and row

Fig. 6 shows the results of making pairwise comparisons for the three groups, using $\varepsilon = 0.3$. The results are in line with the discussion of the differences between the distributions. In the comparisons between group 1 and groups 2 and 3 there are two main regions with important differences in the mass, with group 3 also finding a difference with respect to group 1 in the far tails. The comparison between group 2 and group 3 shows a difference in the left-hand tail but not in the right-hand tail. The fat tails in groups 2 and 3 are generated by the skew Cauchy component of the mixture and the small sample sizes can make this feature difficult to distinguish (in both groups we expect only five observations to come from this distribution).

Changing the value of M^* from 1 to 2 or 3 has very similar effects as in the case of example 1. Using the alternative prior with point masses at zero for M_h leads to virtually identical predictive results and posterior inclusion probabilities of around 0.95 for the common component and the idiosyncratic component of group 1. The component that is shared by groups 2 and 3 has a posterior inclusion probability of around 0.3, whereas, in line with the generating model, the four other components are assigned little posterior mass (less than 0.2).

5.2. Survival analysis

Doss and Huffer (2003) modelled interval-censored data in survival analysis using the Dirichlet process as a prior for the distribution of the survival times. This application focuses on time to cosmetic deterioration of the breast of women with stage 1 breast cancer who have undergone a lumpectomy under two treatments: radiation and radiation with chemotherapy. There are 46 subjects in the radiation-only group and 48 subjects in the combination group. The data have been presented in Beadle *et al.* (1984). The indicator $d_{g,j} = 1$ if the jth person in the gth group suffers an event (in this case retraction of the breast) before the censoring time $T_{g,j}$ and $d_{g,j} = 0$ otherwise. If $d_{g,j} = 1$ then the observation is an interval $A_{g,j}$ in which the event occured. Doss and Huffer (2003) assigned a Dirichlet process prior to the lifetime distribution for each group separately. Since the actual survival times are missing (owing to the interval censoring), the posterior will then be a mixture of Dirichlet processes. Denoting the survival time of individual j in group g by $\tau_{g,j}$, we extend their approach to the model

$$I(\tau_{g,j} \in A_{g,j}) \text{ if } d_{g,j} = 1 \text{ or } I(\tau_{g,j} > T_{g,j}) \text{ if } d_{g,j} = 0,$$

$$\tau_{g,j} \overset{\text{ind}}{\sim} G_g,$$

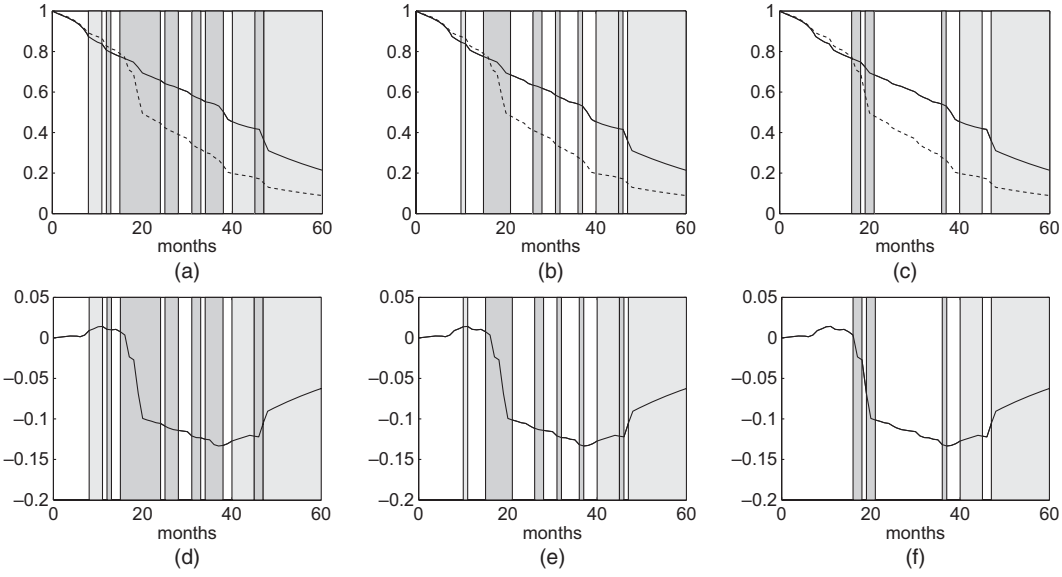

Fig. 7. Survival analysis results (-------, combination group; ———, radiation-only group) showing (a)–(c) the posterior mean survival functions for the two groups and (d)–(f) the posterior mean for Π_1 where the radiation-only group is coded as group 1 (■, more mass for group 2; □, more mass for group 1): (a), (d) $\varepsilon = 0.2$; (b), (e) $\varepsilon = 0.4$; (c), (f) $\varepsilon = 0.6$

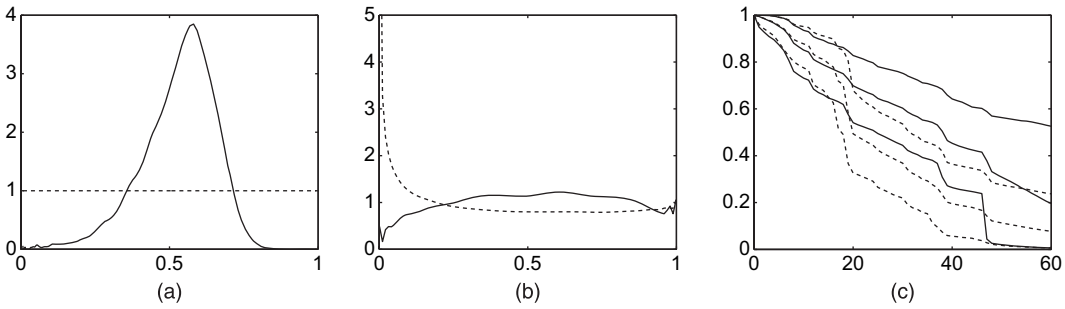

Fig. 8. Survival analysis: prior (-------) and posterior (———) densities of (a) the parameter a and (b) the correlation ρ, and (c) the posterior mean and 95% credible intervals of the survival functions for the combination (-------) and radiation-only (———) groups

$$G_1, G_2, \ldots, G_q \sim \mathrm{CNGG}(M, H, D; a, 1),$$

where H is an exponential distribution with mean $1/\xi$. The parameter ξ is given a vague gamma prior with shape parameter 0.1 and mean 1.

Fig. 7 displays results of the analysis of the clinical trial data. Figs 7(a)–7(c) show that the survival function is similar for the two groups initially but the curves diverge around 16 months with the combination group associated with a much larger number of events. Figs 7(d)–7(f) show the posterior mean of the difference between the survival functions for the groups. They also indicate that the mass is similar until 16 months but then the difference quickly becomes large until the survival functions converge again. The different shades in the graphs also highlight the differences between the distributions for the two groups (rather than the survival functions). The regions that are identified as similar change when moving from $\varepsilon = 0.4$ to $\varepsilon = 0.6$ with the latter having fewer, larger and more connected regions. The results with $\varepsilon = 0.6$ more clearly highlight the larger differences in the survival functions, such as the sharp drop in the combination group around 16 months. Finally, for all values of ε the radiation-only group places more mass than the combination group in the region beyond 45 months.

The posterior distributions of a and ρ are shown in Fig. 8, which indicates that the value $a = 0$ (the Dirichlet process case) is not well supported by the data with a posterior median close to the normalized inverse Gaussian process (where $a = 0.5$), but with substantial posterior uncertainty. The posterior distribution of the correlation parameter ρ indicates that the groups are different but do share some common aspects. This can also be seen in Fig. 8(c), where the credible intervals of the survival functions for both groups are closer together for later survival times than in the independent Dirichlet process analysis of Doss and Huffer (2003). This is in line with the fact that more patients are censored in the radiation group, which shrinks the credible interval towards that for the combination group. The effect of NGG rather than Dirichlet process marginals (larger a induces more small jumps and smoother distributions) is difficult to evaluate on the basis of mean survival functions.

Changing the prior on M_h by varying M^* to 2 and 3 has no discernible effect on the predictive distributions and very little effect on the posterior for a and ρ. Posterior distributions for M_h are somewhat affected by this, with posterior medians increasing by a factor of about 2 as M^* is increased from 1 to 3.

If we use the point mass prior on M_h, we find again that the predictive distributions are virtually unchanged, and the common component is included with posterior probability 1, whereas the posterior inclusion probabilities of the idiosyncratic components are about 0.6 for group 1 and 0.2 for group 2.

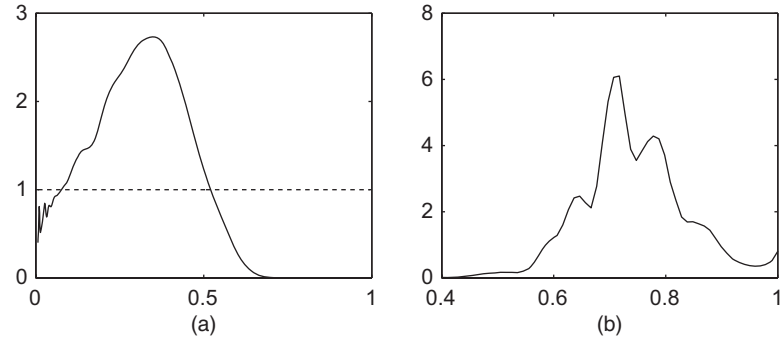

Fig. 9. Stochastic frontier analysis: (a) posterior (———) and prior (- - - - - - -) densities of a and (b) the density of the posterior mean of the average efficiency distribution with an NGG prior

5.3. Stochastic frontier analysis

Stochastic frontier analysis is a popular method in econometrics for estimating the efficiency of firms. We shall consider an application to the efficiency of US hospitals by using data that have previously been analysed by Koop *et al.* (1997). It is assumed that all hospitals operate relative to a common cost frontier, which represents the minimum cost of performing the functions of that hospital (including operations, patient care, etc.). Inefficiency can then be measured by how far a hospital operates above the optimal cost level that is given by the frontier. The costs are observed for the hospitals over a number of years. The model is written in terms of log-cost $C_{g,j,t}$ for the jth hospital in the gth group at the tth time point

$$C_{g,j,t} = \alpha + x_{g,j,t}^{\mathrm{T}}\beta + u_{g,j} + \varepsilon_{g,j,t},$$

where $x_{g,j,t}$ are variables used to define the frontier for hospital j in group g at time t, $u_{g,j} > 0$ is the inefficiency for hospital j in group g and $\varepsilon_{g,j,t}$ are mutually independent measurement errors which will be assumed to be normally distributed with mean 0 and variance σ^2. The model assumes that the efficiency of hospitals is constant over the observed time period (which is a common assumption in the applied literature). The efficiency for hospital j in group g is defined to be $\exp(-u_{g,j})$.

The main focus of this type of analysis is the estimation of the hospital efficiencies $\exp(-u_{g,j})$. A Bayesian non-parametric analysis of the stochastic frontier model is described by Griffin and Steel (2004) who assumed a Dirichlet process prior for the inefficiency distribution and used the same data set. The model used in the present paper is

$$C_{g,j,t} \overset{\text{ind}}{\sim} N(\alpha + x_{g,j,t}^{\mathrm{T}}\beta + u_{g,j}, \sigma^2),$$

$$u_{g,j} \overset{\text{ind}}{\sim} G_g$$

$$G_1, G_2, \ldots, G_q \sim \mathrm{CNGG}(M, H, D; a, 1),$$

where α, β and σ^2 are given the priors that were described by Griffin and Steel (2004) and H is an exponential distribution with mean $1/\xi$, where ξ is given an exponential prior with mean $-1/\log(r^*)$, so that r^* is the prior median efficiency. In this example r^* is subjectively chosen to take the value 0.8. Fig. 9 shows some posterior results of extending the model of Griffin and Steel (2004) using the prior that was developed in this paper for just one group. The posterior distribution of a has a mode at around 0.4. The density of the posterior mean of the efficiency

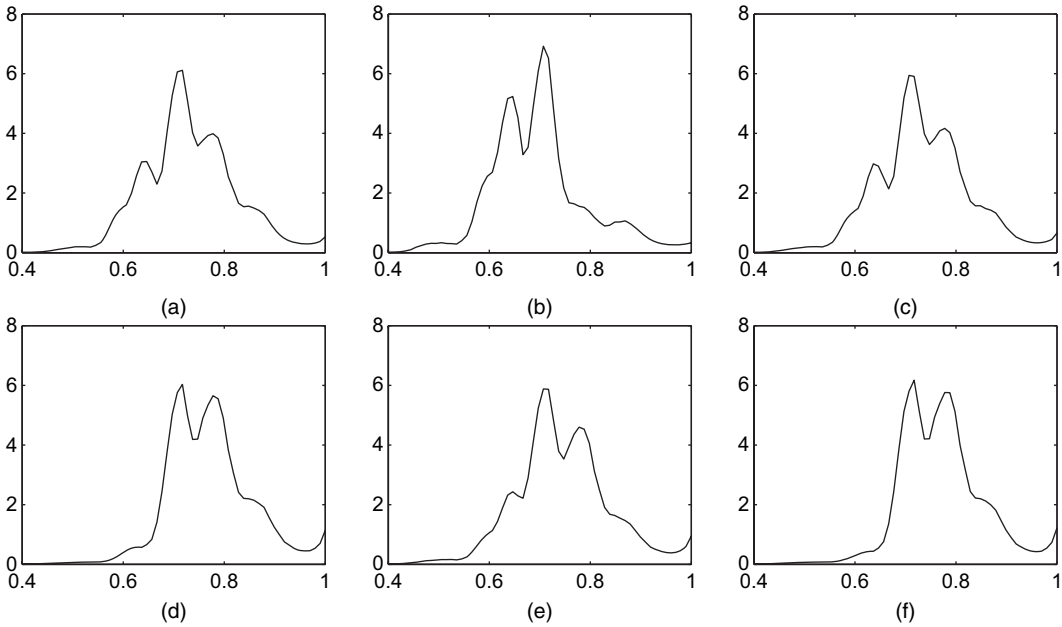

Fig. 10. Stochastic frontier analysis—density of the posterior mean of the efficiency distribution for each hospital type with a CNGG prior: (a) for profit, low SPR; (b) non-profit, low SPR; (c) government, low SPR; (d) for profit, high SPR; (e) non-profit, high SPR; (f) government, high SPR

distribution with just a single group (averaged over all hospital types) has three internal modes at roughly 0.65, 0.7 and 0.8 and a further mode at 1, which is quite in line with the results for the efficiency that were obtained in Griffin and Steel (2004) based on all the data. We now use information about the type of hospital expressed by two discrete covariates: the ownership status of the hospital (for profit, non-profit and government) and a quality factor in terms of the staff–patient ratio SPR (low or high). The precise definition of these covariates is described in Koop *et al.* (1997).

Fig. 10 shows the density corresponding to the posterior mean of the efficiency distribution within each group defined by these two covariates. It should be noted that these densities exist since the kernel is continuous in $u_{g,j}$ and h is continuous, which implies that the posterior means of $G_1, \ldots, G_p$ are continuous. For comparison, an analysis using a product of Dirichlet processes was provided by Griffin and Steel (2004). They specified separate non-parametric inefficiency distributions for each group, which are linked only indirectly by the centring distribution, which is allowed to depend on firm characteristics, and a common mass parameter. The dependent non-parametric prior that was developed in the present paper leads to predictive distributions which vary substantially less between groups, illustrating the model's ability to borrow information effectively. This is particularly important in this application where group sizes are quite small, ranging from 20 to 141. All distributions are multimodal with most densities having modes at roughly 0.7 and 0.8 (and at 1). However, the sizes of the modes differ between the groups.

The decomposition in equation (8) more clearly shows the differences and similarities between the distributions. Fig. 11 shows the density $\pi_{i\cdot}$ of the posterior mean of $\Pi_{i\cdot}$, $\pi_{\cdot j}$ of the posterior mean of $\Pi_{\cdot j}$, and $\gamma_{i,j}$ of the posterior mean of $\Gamma_{i,j}$ (again, these densities exist). The πs show the effect of one factor averaging over the other factors. Hospitals with high SPR tend to have more mass at higher efficiency than low SPR hospitals (suggesting that they tend to be more

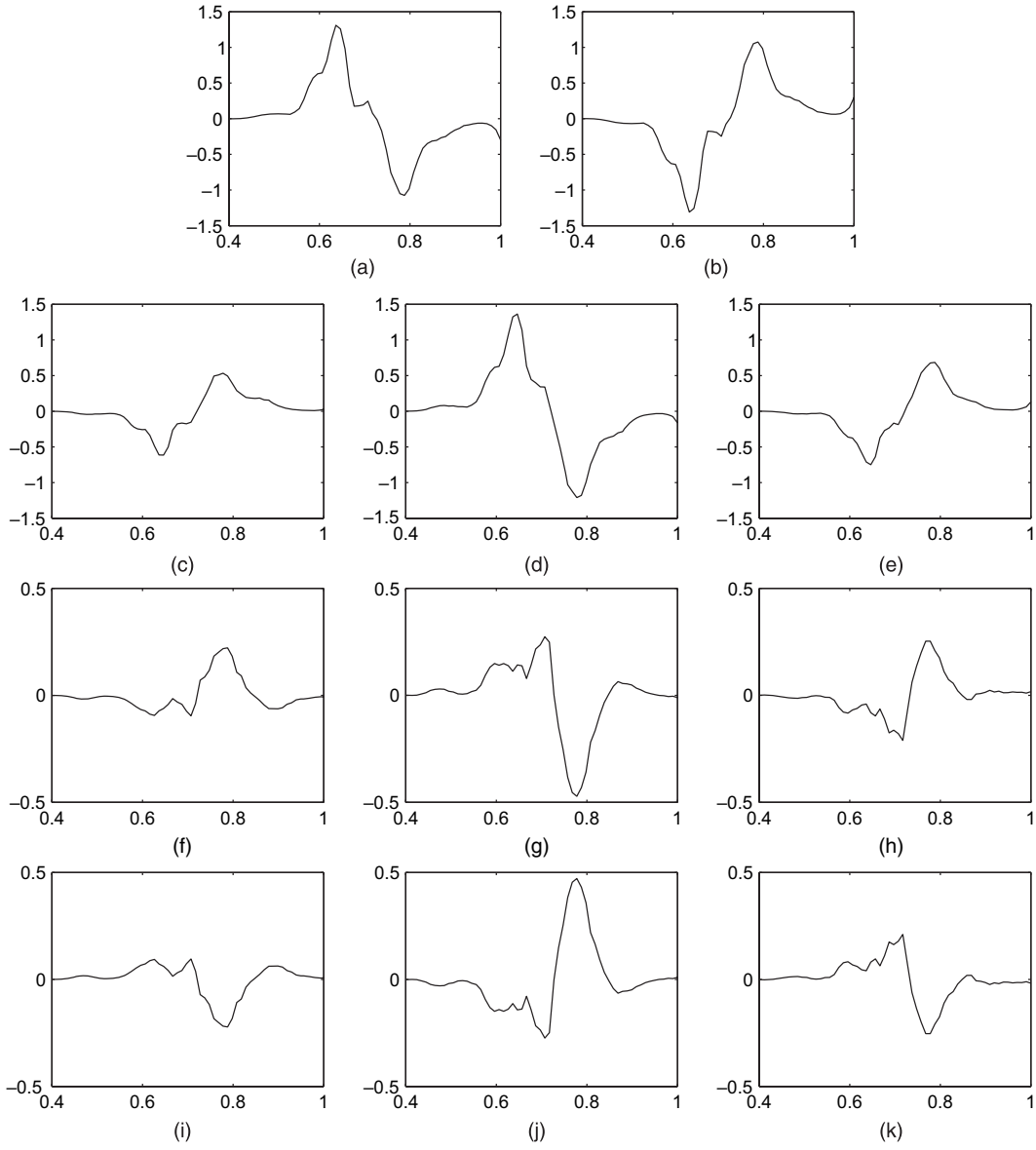

Fig. 11. Stochastic frontier analysis—posterior means of (a)–(e) $\pi_{i\cdot}$ and $\pi_{\cdot j}$, and (f)–(k) $\gamma_{i,j}$ with NGG process marginals: (a) low SPR; (b) high SPR; (c) for-profit hospital; (d) non-profit hospital; (e) government hospital; (f) for profit, low SPR; (g) non-profit, low SPR; (h) government, low SPR; (i) for profit, high SPR; (j) non-profit, high SPR; (k) government, high SPR

efficient). The effect of high SPR is to shift mass away from around 0.65 to around 0.8. The for-profit and government hospitals have similar distributions and have more mass at higher level efficiency than non-profit hospitals do, again mostly involving shifts from regions around 0.65 to those near 0.8. The densities γ relate to interaction terms which are most important for non-profit hospitals where non-profit hospitals with low SPR tend to have particularly low mass at high levels of efficiency (around 0.8). Thus, the results clearly indicate which factors (or combinations of factors) lead to distributions that place more mass on higher levels of efficiency.

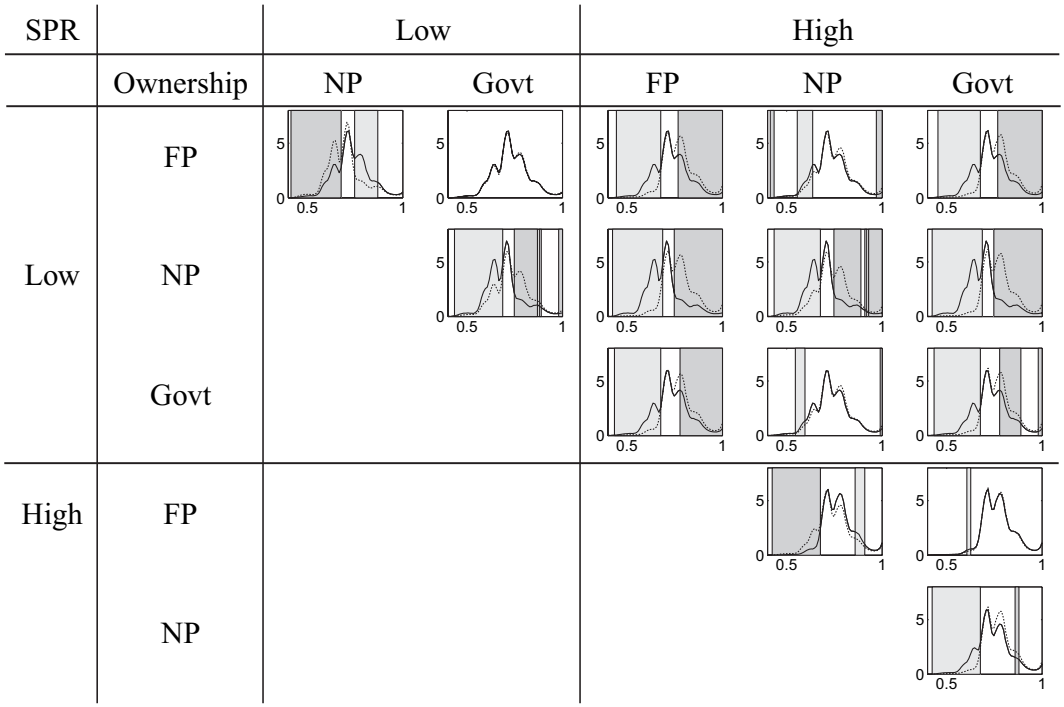

Fig. 12. Stochastic frontier analysis—pairwise comparisons of efficiency distributions according to type of ownership (FP, for profit; NP, non-profit; Govt, government) and SPR: the pairs are shown as the row (———) and column (-------), with dark grey shading indicating higher mass in the column and light grey shading indicating higher mass in the row

Fig. 12 shows pairwise comparisons of the distributions which identify regions where the mass placed by the two corresponding distributions is substantially different, using $\varepsilon = 0.3$. Simply considering the overplotted densities it is clear that there is very little difference between the for-profit and government hospitals at both levels of quality (in line with their very similar πs). There is also little difference between the non-profit hospitals at high quality and the for-profit and government hospital at low quality (the πs for both factors virtually balance each other out and the γs are similar). The other combinations of factors lead to clear results where we can identify regions of the support where one distribution places more mass than the other and vice versa. Clearly, the for-profit and government hospitals with high quality are the most efficient combinations, placing more mass on higher efficiencies than other cases. Interestingly, the much more restrictive fully parametric model without interactions of Koop *et al.* (1997) leads to the very different (and counterintuitive) conclusions that for-profit status and high SPR both reduce efficiencies.

Using values of M^* equal to 2 and 3 rather than $M^* = 1$ makes no real difference to the results presented above, with the possible exception of the posterior means of $\gamma_{i,j}$ which are somewhat affected. Posterior medians of M_h are also moderately affected by varying the choice of M^*, changing by a factor of up to 2. The adoption of prior point masses at zero for M_h has virtually no effect on the reported results for a and the group efficiency distributions and a slight effect on the posterior mean of the πs and does lead to some changes for the γs. Pairwise comparisons are quite similar to those reported in Fig. 12, with minor differences which do not

change any of the conclusions. Posterior inclusion probabilities of two components are near 1: the component shared by all low SPR groups and non-profit hospitals with high SPR, and the component shared by all high SPR hospitals as well as government hospitals with low SPR. Both of these sets of groups are quite homogeneous with relatively little difference between the efficiency distributions (if we use $\varepsilon = 0.4$ instead, the comparison in Fig. 12 identifies very few differences within these sets). All 61 other components are assigned relatively little posterior mass (two have in between 0.1 and 0.2, one receives 0.09 and the rest receive 0.05 or less).

6. Summarizing remarks

This paper proposes a methodology for inferring differences between distributions that are associated with different groups of observations. A Bayesian non-parametric approach is taken and we introduce a novel form of prior, derived from NRMIs. The prior allows the inclusion of information about partial exchangeability and so represents prior beliefs which could not be expressed by using for example the hierarchical Dirichlet process. This leads to effective borrowing of strength between distributions without assuming exchangeability and can easily and systematically accommodate widely varying levels of complexity in terms of dependence. Efficient, exact inference is possible by using a slice sampling method, which extends the ideas of Griffin and Walker (2011). The prior is used with a new graphical method to compare pairs of distributions. The common support of any two distributions is partitioned and each element of the partition is characterized by obtaining more mass from either distribution or being allocated roughly similar mass by both distributions. We believe that this is an effective way of understanding and communicating the difference between distributions. In particular, where the groups are defined by several covariates, we propose an informative ANOVA-type decomposition of the differences.

We analyse applications in survival analysis and stochastic frontiers with small numbers of observations, which are typical of real data applications in many fields. Despite this, the models perform very well and lead to sensible results. Interestingly, in both applications, models with Dirichlet process marginals are not well supported by the data and NGG marginals are favoured. The posterior distribution of a in the survival example is centred on 0.5, which corresponds to the normalized inverse Gaussian process. We have used two prior structures on the mass parameters M_h: one prior which encourages shrinkage towards zero and one which incorporates prior point masses at zero. Changing the prior assumptions on M_h does not have a substantive effect on the conclusions in our applications.

We believe that the methodology proposed in this paper is highly flexible, yet widely applicable to real data, and allows for quite informative inference on the (sources of the) differences between dependent distributions.

Acknowledgements

We are very grateful to two referees and the Associate Editor for thoughtful and constructive comments. Mark Steel gratefully acknowledges the hospitality of the Universidad Carlos III de Madrid during the latter stages of this research. Michalis Kolossiatis was partially funded by the Cyprus University of Technology.

Appendix A: Proofs

A.1. Proof of theorem 1

A similar approach is taken to James *et al.* (2006, 2009). We know that $E[G_1(B)] = E[G_2(B)] = H(B)$. To

calculate the covariance, we need

$$\begin{aligned}E[G_1(B)\,G_2(B)] &= E\left[\frac{\tilde{G}_1(B)}{\tilde{G}_1(\Omega)}\frac{\tilde{G}_2(B)}{\tilde{G}_2(\Omega)}\right]\\ &= E\left[\frac{\{\tilde{G}_1^*(B)+\tilde{G}_3^*(B)\}\{\tilde{G}_2^*(B)+\tilde{G}_3^*(B)\}}{\{\tilde{G}_1^*(\Omega)+\tilde{G}_3^*(\Omega)\}\{\tilde{G}_2^*(\Omega)+\tilde{G}_3^*(\Omega)\}}\right]\\ &= \int_0^\infty\int_0^\infty E\left[\gamma(v_1,v_2)\right]\mathrm{d}v_1\,\mathrm{d}v_2,\end{aligned}$$

where

$$\begin{aligned}\gamma(v_1,v_2) &= \{\tilde{G}_1^*(B)+\tilde{G}_3^*(B)\}\{\tilde{G}_2^*(B)+\tilde{G}_3^*(B)\}\exp[-v_1\{\tilde{G}_1^*(\Omega)+\tilde{G}_3^*(\Omega)\}-v_2\{\tilde{G}_2^*(\Omega)+\tilde{G}_3^*(\Omega)\}]\\ &= \{\tilde{G}_1^*(B)\,\tilde{G}_2^*(B)+\tilde{G}_3^*(B)\tilde{G}_2^*(B)+\tilde{G}_1^*(B)\tilde{G}_3^*(B)+\tilde{G}_3^*(B)^2\}\\ &\quad\times\exp\{-v_1\,\tilde{G}_1^*(\Omega)-v_2\,\tilde{G}_2^*(\Omega)-(v_1+v_2)\,\tilde{G}_3^*(\Omega)\}.\end{aligned}$$

The independence of the underlying processes $\tilde{G}_1^*, \tilde{G}_2^*$ and $\tilde{G}_3^*$ and the independence of Lévy processes on disjoint sets gives

$$\begin{aligned}E[\gamma(v_1,v_2)] &= E[\tilde{G}_3^*(B)^2\exp\{-(v_1+v_2)\,\tilde{G}_3^*(B)\}]E[\exp\{-(v_1+v_2)\,\tilde{G}_3^*(B^c)\}]E[\exp\{-v_1\,\tilde{G}_1^*(\Omega)\}]\\ &\quad\times E[\exp\{-v_2\,\tilde{G}_2^*(\Omega)\}]+E[\tilde{G}_3^*(B)\exp\{-(v_1+v_2)\,\tilde{G}_3^*(B)\}]E[\tilde{G}_2^*(B)\exp\{-v_2\,\tilde{G}_2^*(B)\}]\\ &\quad\times E[\exp\{-(v_1+v_2)\,\tilde{G}_3^*(B^c)\}]E[\exp\{-v_1\,\tilde{G}_1^*(\Omega)\}]E[\exp\{-v_2\,\tilde{G}_2^*(B^c)\}]\\ &\quad+E[\tilde{G}_1^*(B)\exp\{-v_1\,\tilde{G}_1^*(B)\}]E[\tilde{G}_3^*(B)\exp\{-(v_1+v_2)\,\tilde{G}_3^*(B)\}]E[\exp\{-(v_1+v_2)\,\tilde{G}_3^*(B^c)\}]\\ &\quad\times E[\exp\{-v_1\,\tilde{G}_1^*(B^c)\}]E[\exp\{-v_2\,\tilde{G}_2^*(\Omega)\}]+E[\tilde{G}_1^*(B)\exp\{-v_1\,\tilde{G}_1^*(B)\}]\\ &\quad\times E[\tilde{G}_2^*(B)\exp\{-v_2\,\tilde{G}_2^*(B)\}]E[\exp\{-(v_1+v_2)\,\tilde{G}_3^*(\Omega)\}]E[\exp\{-v_1\,\tilde{G}_1^*(B^c)\}]\\ &\quad\times E[\exp\{-v_2\,\tilde{G}_2^*(B^c)\}]\end{aligned}$$

The definition of $L_\eta(v)$ implies that $E[\exp\{-v\,\tilde{G}_k^*(B)\}]=\exp\{-H(B)M_k\,L_\eta(v)\}$ and then

$$\begin{aligned}E[\tilde{G}_k^*(B)\exp\{-v\,\tilde{G}_k^*(B)\}] &= -E\left[\frac{\mathrm{d}}{\mathrm{d}v}\exp\{-v\,\tilde{G}_k^*(B)\}\right]=-\frac{\mathrm{d}}{\mathrm{d}v}E[\exp\{-v\,\tilde{G}_k^*(B)\}]\\ &= -\frac{\mathrm{d}}{\mathrm{d}v}\exp\{-H(B)M_k\,L_\eta(v)\}=H(B)M_k\,L_\eta'(v)\exp\{-H(B)M_k\,L_\eta(v)\},\end{aligned}$$

$$\begin{aligned}E[\tilde{G}_k^*(B)^2\exp\{-v\,\tilde{G}_k^*(B)\}] &= E\left[\frac{\mathrm{d}^2}{\mathrm{d}v^2}\exp\{-v\,\tilde{G}_k^*(B)\}\right]=\frac{\mathrm{d}^2}{\mathrm{d}v^2}E[\exp\{-v\,\tilde{G}_k^*(B)\}]\\ &= [H(B)^2M_k^2L_\eta'(v)^2-H(B)M_k\,L_\eta''(v)]\exp\{-H(B)M_k\,L_\eta(v)\}.\end{aligned}$$

It follows that

$$\begin{aligned}E[\gamma(v_1,v_2)] &= \{H(B)^2M_3^2\,L_\eta'(v_1+v_2)^2-H(B)M_3\,L_\eta''(v_1+v_2)\}\exp\{-H(B)M_3\,L_\eta(v_1+v_2)\}\\ &\quad\times\exp[-\{1-H(B)\}M_3\,L_\eta(v_1+v_2)]\exp\{-M_1\,L_\eta(v_1)\}\exp\{-M_2\,L_\eta(v_2)\}\\ &\quad+H(B)M_3\,L_\eta'(v_1+v_2)\exp\{-H(B)M_3\,L_\eta(v_1+v_2)\}H(B)M_2\,L_\eta'(v_2)\exp\{-H(B)M_2\,L_\eta(v_2)\}\\ &\quad\times\exp[-\{1-H(B)\}M_3\,L_\eta(v_1+v_2)]\exp\{-M_1\,L_\eta(v_1)\}\exp[-\{1-H(B)\}M_2\,L_\eta(v_2)]\\ &\quad+H(B)M_1\,L_\eta'(v_1)\exp\{-H(B)M_1\,L_\eta(v_1)\}\,H(B)M_3\,L_\eta'(v_1+v_2)\exp\{-H(B)M_3\,L_\eta(v_1+v_2)\}\\ &\quad\times\exp[-\{1-H(B)\}M_3\,L_\eta(v_1+v_2)]\exp[-\{1-H(B)\}M_1\,L_\eta(v_1)]\exp\{-M_2\,L_\eta(v_2)\}\\ &\quad+H(B)M_1\,L_\eta'(v_1)\exp\{-H(B)M_1\,L_\eta(v_1)\}\,H(B)M_2\,L_\eta'(v_2)\exp\{-H(B)M_2\,L_\eta(v_2)\}\\ &\quad\times\exp\{-M_3\,L_\eta(v_1+v_2)\}\exp[-\{1-H(B)\}M_1\,L_\eta(v_1)]\exp[-\{1-H(B)\}\,M_2\,L_\eta(v_2)]\\ &= [H(B)^2\{M_1\,L_\eta'(v_1)+M_3\,L_\eta'(v_1+v_2)\}\{M_2\,L_\eta'(v_2)+M_3\,L_\eta'(v_1+v_2)\}-H(B)\,M_3\,L_\eta''(v_1+v_2)]\\ &\quad\times\exp\{-M_3\,L_\eta(v_1+v_2)\}\exp\{-M_1\,L_\eta(v_1)\}\exp\{-M_2\,L_\eta(v_2)\}.\end{aligned}$$

Then

$$\operatorname{cov}\{G_1(B), G_2(B)\} = H(B)^2 \left\{ \int_0^\infty \int_0^\infty \alpha\gamma \, \mathrm{d}v_1 \, \mathrm{d}v_2 - 1 \right\} - H(B) \int_0^\infty \int_0^\infty \beta\gamma \, \mathrm{d}v_1 \, \mathrm{d}v_2,$$

where $\alpha = \{M_1 L'_\eta(v_1) + M_3 L'_\eta(v_1 + v_2)\}\{M_2 L'_\eta(v_2) + M_3 L'_\eta(v_1 + v_2)\}$, $\beta = M_3 L''_\eta(v_1 + v_2)$ and $\gamma = \exp\{-M_3 L_\eta(v_1 + v_2) - M_1 L_\eta(v_1) - M_2 L_\eta(v_2)\}$. The result follows from

$$\int_0^\infty \int_0^\infty \alpha\gamma \, \mathrm{d}v_1 \, \mathrm{d}v_2 = 1 + \int_0^\infty \int_0^\infty M_3 L''_\eta(v_1 + v_2) \exp\{-M_1 L_\eta(v_1) - M_2 L_\eta(v_2) - M_3 L_\eta(v_1 + v_2)\} \, \mathrm{d}v_1 \, \mathrm{d}v_2.$$

A.2. Proof of theorem 2

Since H is non-atomic, every distinct value is associated with one, and only one, jump. Suppose that the z_jth underlying unnormalized random measure (so that $1 \leqslant z_j \leqslant p$) generated the jth distinct value. Let K_l^* be the number of distinct values for which $z_j = l$ (the notation in these proofs does not make the dependence on z explicit). We let the ith distinct value be the a_ith distinct value in the z_ith underlying unnormalized random measure. These values are defined so that $\{a_i | z_i = j\}$ takes values in $\{1, 2, \ldots, K_j^*\}$ and $a_i = a_j$ if $z_i = z_j$ only if $i = j$ (so that the values of a_i are distinct). It is also convenient to write $m_{z_i, a_i} = m_i$. Let $p_{j,i}$ be the probability of the ith distinct value in the jth group; then the partition probability function for known $z_1, \ldots, z_K$ is

$$\begin{aligned}
E\left[\prod_{i=1}^{K} \prod_{g=1}^{q} p_{g,i}^{n_{g,i}} \right] &= E\left[\prod_{i=1}^{K} \prod_{g=1}^{q} \left(\frac{J_{z_i, a_i}}{\sum_{j=1}^{p} \sum_{k=1}^{\infty} D_{gj} J_{j,k}} \right)^{n_{g,i}} \right] \\
&= E\left[\frac{\prod_{i=1}^{K} J_{z_i, a_i}^{m_i}}{\prod_{g=1}^{q} \left(\sum_{j=1}^{p} \sum_{k=1}^{\infty} D_{gj} J_{j,k} \right)^{n_g}} \right] \\
&= E\left[\frac{\prod_{j=1}^{p} \prod_{i=1}^{K_j^*} J_{j,i}^{m_{j,i}}}{\prod_{g=1}^{q} \left(\sum_{j=1}^{p} \sum_{k=1}^{\infty} D_{gj} J_{j,k} \right)^{n_g}} \right].
\end{aligned}$$

As in James *et al.* (2006) the identity

$$\frac{1}{y^n} = \frac{1}{\Gamma(n)} \int v^{n-1} \exp(-vy) \, \mathrm{d}v$$

can be repeatedly used on this equation to give

$$\begin{aligned}
&E\left[\int_{(0,\infty)^q} \prod_{g=1}^{q} \frac{1}{\Gamma(n_g)} v_g^{n_g - 1} \prod_{j=1}^{p} \prod_{i=1}^{K_j^*} J_{j,i}^{m_{j,i}} \exp\left(- \sum_{g=1}^{q} D_{gj} v_g \sum_{k=1}^{\infty} J_{j,k} \right) \mathrm{d}v \right] \\
&\qquad = \int_{(0,\infty)^q} \prod_{g=1}^{q} \frac{1}{\Gamma(n_g)} v_g^{n_g - 1} \prod_{j=1}^{p} E\left[\prod_{i=1}^{K_j^*} J_{j,i}^{m_{j,i}} \exp\left(- \sum_{g=1}^{q} D_{gj} v_g \sum_{k=1}^{\infty} J_{j,k} \right) \right] \mathrm{d}v.
\end{aligned}$$

Applying the set limiting argument of James *et al.* (2006), appendix A, to this equation leads to

$$\begin{aligned}
\int_{(0,\infty)^q} \prod_{g=1}^{q} \frac{1}{\Gamma(n_g)} v_g^{n_g - 1} \prod_{j=1}^{p} M_j^{K_j^*} \prod_{i=1}^{K_j^*} \int J_{j,i}^{m_{j,i}} \exp\left(- \sum_{g=1}^{q} D_{gj} v_g J_{j,i} \right) \eta(J_{j,i}) \, \mathrm{d}J_{j,i} \\
\times \prod_{j=1}^{p} E\left[\exp\left(- \sum_{g=1}^{q} D_{gj} v_g \sum_{k=1}^{\infty} J_{j,k} \right) \right] \mathrm{d}v
\end{aligned}$$

$$= \int_{(0,\infty)^q} \prod_{g=1}^{q} \frac{1}{\Gamma(n_g)} v_g^{n_g-1} \prod_{j=1}^{p} M_j^{K_j^*} \prod_{i=1}^{K_j^*} I_\eta\left(m_{j,i}, \sum_{g=1}^{q} D_{gj} v_g\right) \prod_{j=1}^{p} \exp\left\{-M_j L_\eta\left(\sum_{g=1}^{q} D_{gj} v_g\right)\right\} \mathrm{d}v$$

$$= \int_{(0,\infty)^q} \prod_{g=1}^{q} \frac{1}{\Gamma(n_g)} v_g^{n_g-1} \prod_{j=1}^{p} M_j^{K_j^*} \prod_{i=1}^{K} I_\eta\left(m_i, \sum_{g=1}^{q} D_{gz_i} v_g\right) \prod_{j=1}^{p} \exp\left\{-M_j L_\eta\left(\sum_{g=1}^{q} D_{gj} v_g\right)\right\} \mathrm{d}v,$$

where

$$I_\eta(n, v) = \int J^n \exp(-vJ)\, \eta(J)\, \mathrm{d}J.$$

The value of z_j is unknown. However, z_j can only take the value k if the kth column of D contains a 0 in every row i for which $n_{i,j} = 0$. Therefore the set of possible values for z_j is given by the set a_j. The result arises from summing over the possible values of $z_1, \ldots, z_K$.

A.3. Proof of corollary 1

If the CDP model is assumed then $\eta(x) = x^{-1}\exp(-x)$ and so $L_\eta(v) = \log(1+v)$ and

$$I_\eta(n, v) = \int J^{n-1} \exp\{-(1+v)J\}\, \mathrm{d}J = \Gamma(n)(1+v)^{-n}.$$

Theorem 2 implies that the partition probability function is

$$\sum_{z \in \mathcal{Z}} \int_{(0,\infty)^q} \frac{\prod_{i=1}^{K} \Gamma(m_i)}{\prod_{g=1}^{q} \Gamma(n_g)} \prod_{g=1}^{q} v_g^{n_g-1} \prod_{j=1}^{p} M_j^{K_j^*} \left(1 + \sum_{g=1}^{q} D_{gj} v_g\right)^{-M_j^*} \mathrm{d}v.$$

If $q = 2$, $p = 3$, and

$$D = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix},$$

then the partition probability function is

$$\sum_{z \in \mathcal{Z}} \int_0^\infty \int_0^\infty \frac{\prod_{i=1}^{K} \Gamma(m_i)}{\prod_{g=1}^{2} \Gamma(n_g)} \prod_{j=1}^{3} M_j^{K_j^*} v_1^{n_1-1} v_2^{n_2-1} (1+v_1+v_2)^{-M_3^*} (1+v_1)^{-M_1^*} (1+v_2)^{-M_2^*}\, \mathrm{d}v_1\, \mathrm{d}v_2.$$

Reparameterizing to $x_1 = v_1/(1+v_1)$ and $x_2 = v_2/(1+v_2)$ gives

$$\sum_{z \in \mathcal{Z}} \int_0^1 \int_0^1 \frac{\prod_{i=1}^{K} \Gamma(m_i)}{\prod_{g=1}^{2} \Gamma(n_g)} \prod_{j=1}^{3} M_j^{K_j^*} x_1^{n_1-1} x_2^{n_2-1} (1-x_1)^{M_1^*+M_3^*-n_1-1} (1-x_2)^{M_2^*+M_3^*-n_2-1} (1-x_1x_2)^{-M_3^*}\, \mathrm{d}x_1 \mathrm{d}x_2$$

$$= \sum_{z \in \mathcal{Z}} \frac{\Gamma(M_1^*+M_3^*-n_1)}{\Gamma(M_1^*+M_3^*)} \frac{\prod_{i=1}^{K} \Gamma(m_i)}{\Gamma(n_2)} \prod_{j=1}^{3} M_j^{K_j^*} \int_0^\infty x_2^{n_2-1} (1-x_2)_2^{M_2^*+M_3^*-n_2-1} {}_2F_1(M_3^*, n_1; M_1^*+M_3^*; x_2)\, \mathrm{d}x_2$$

$$= \sum_{z \in \mathcal{Z}} \prod_{i=1}^{K} \Gamma(m_i) \prod_{j=1}^{3} M_j^{K_j^*} \frac{\Gamma(M_1^*+M_3^*-n_1)}{\Gamma(M_1^*+M_3^*)} \frac{\Gamma(M_2^*+M_3^*-n_2)}{\Gamma(M_2^*+M_3^*)} {}_3F_2(M_3^*, n_1, n_2; M_1^*+M_3^*, M_2^*+M_3^*; 1),$$

where ${}_qF_p$ is the generalized hypergeometric function.

References

Beadle, G., Come, S., Henderson, C., Silver, B. and Hellman, S. (1984) The effect of adjuvant chemotherapy on the cosmetic results after primary radiation treatment for early stage breast cancer. *Int. J. Radn Oncol. Biol. Phys.*, **10**, 2131–2137.

Brix, A. (1999) Generalized gamma measures and shot-noise Cox processes. *Adv. Appl. Probab.*, **31**, 929–953.
De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004) An ANOVA model for dependent random measures. *J. Am. Statist. Ass.*, **99**, 205–215.
Doss, H. and Huffer, F. W. (2003) Monte Carlo methods for Bayesian analysis of survival data using mixtures of Dirichlet process prior. *J. Computnl Graph. Statist.*, **12**, 282–307.
Dunson, D. B., Xue, Y. and Carin, L. (2008) The matrix stick breaking process: flexible Bayes meta analysis. *J. Am. Statist. Ass.*, **103**, 317–327.
Ferguson, T. S. (1973) A Bayesian analysis of some non-parametric problems. *Ann. Statist.*, **1**, 209–230.
Fernández, C. and Steel, M. F. J. (1998) On Bayesian modeling of fat tails and skewness. *J. Am. Statist. Ass.*, **93**, 359–371.
Griffin, J. E. (2011) The Ornstein-Uhlenbeck Dirichlet process and other time-varying processes for Bayesian non-parametric inference. *J. Statist. Planng Inf.*, **141**, 3648–3664.
Griffin, J. E. and Brown, P. J. (2010) Inference with Normal-Gamma prior distributions in regression problems. *Baysn Anal.*, **5**, 171–188.
Griffin, J. E. and Steel, M. F. J. (2004) Semiparametric Bayesian inference for stochastic frontier models. *J. Econmetr.*, **123**, 121–152.
Griffin, J. E. and Walker, S. G. (2011) Posterior simulation of Normalised Random Measure mixtures. *J. Computnl Graph. Statist.*, **20**, 241–259.
James, L. F., Lijoi, A. and Prünster, I. (2006) Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. Statist.*, **33**, 105–120.
James, L. F., Lijoi, A. and Prünster, I. (2009) Posterior analysis for normalized random measures with independent increments. *Scand. J. Statist.*, **36**, 76–97.
Kalli, M., Griffin, J. E. and Walker, S. G. (2011) Slice sampling mixture models. *Statist. Comput.*, **21**, 93–105.
Kingman, J. F. C. (1975) Random discrete distributions (with discussion). *J. R. Statist. Soc.* B, **37**, 1–22.
Kolossiatis, M., Griffin, J. E. and Steel, M. F. J. (2012) On Bayesian nonparametric modelling of two correlated distributions. *Statist. Comput.*, to be published.
Koop, G., Osiewalski, J. and Steel, M. F. J. (1997) Bayesian efficiency analysis through individual effects: hospital cost frontiers. *J. Econmetr.*, **76**, 77–105.
Lijoi, A., Mena, R. H. and Prünster, I. (2005) Hierarchical mixture modeling with normalized inverse-Gaussian priors. *J. Am. Statist. Ass.*, **100**, 1278–1291.
Lijoi, A., Mena, R. H. and Prünster, I. (2007) Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Statist. Soc.* B, **69**, 715–740.
Lijoi, A., Nipoti, B. and Prünster, I. (2011) Bayesian inference with dependent normalized completely random measures. *Working Paper 224.* Collegio Carlo Alberto, Moncalieri.
Müller, P., Quintana, F. and Rosner, G. (2004) A method for combining inference across related non-parametric Bayesian models. *J. R. Statist. Soc.* B, **66**, 735–749.
Nieto-Barajas, L. E. and Prünster, I. (2009) A sensitivity analysis of Bayesian non-parametric density estimators. *Statist. Sin.*, **19**, 685–705.
Rodriguez, A., Dunson, D. and Taylor, J. (2009) Bayesian hierarchically weighted finite mixture models for samples of distributions. *Biostatistics*, **10**, 155–171.
Scott, J. G. and Polson, N. G. (2011) Shrink globally, act locally: sparse Bayesian regularization and prediction (with discussion). In *Bayesian Statistics 9* (eds J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West), pp. 501–538. Oxford: Oxford University Press.
Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006) Hierarchical Dirichlet processes. *J. Am. Statist. Ass.*, **101**, 1566–1581.
Walker, S. G. (2007) Sampling the Dirichlet mixture model with slices. *Communs Statist. Simuln Computn*, **36**, 45–54.